Differenzielle Validität von Mathematiktestaufgaben für Kinder mit nicht-deutscher Familiensprache – Welche Rolle spielt die sprachliche Komplexität der Aufgaben?

Dissertation
Zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)
im Fach Psychologie

eingereicht an der
Lebenswissenschaftlichen Fakultät der
Humboldt-Universität zu Berlin

von Dipl.-Psych. Nicole Haag

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/Gutachterin:

1. Prof. Dr. Oliver Lüdtke
2. Prof. Dr. Petra Stanat
3. Prof. Dr. Martin Brunner

Tag der Verteidigung: 18. Dezember 2015

Inhalt

Danksagung

Die hier vorliegende Arbeit entstand während meiner Zeit als wissenschaftliche Mitarbeiterin am Institut zur Qualitätsentwicklung im Bildungswesen. Ohne die Unterstützung zahlreicher Personen wäre die Anfertigung dieser Arbeit nicht möglich gewesen.

Mein besonderer Dank gilt Prof. Dr. Petra Stanat für die Betreuung dieser Arbeit. Ihre inhaltliche Expertise und ihre ausführlichen und konstruktiven Anmerkungen sowie ihre Bereitschaft, meine Manuskripte immer wieder zu lesen und zu kommentieren, haben wesentlich zur vorliegenden Arbeit beigetragen. Ebenso möchte ich mich bei Birgit Heppt für die wunderbare Zusammenarbeit, die zahlreichen und fruchtbaren fachlichen Diskussionen und insbesondere für ihren unermüdlichen Optimismus bedanken. Eine bessere Koautorin als dich hätte ich mir nicht wünschen können. Außerdem danke ich Alexander Roppelt für seine methodische Expertise und für das Interesse, das er meiner Arbeit über die Jahre hinweg entgegen gebracht hat. Allen an der *International Max Planck Research School LIFE* beteiligten Personen bin ich dafür dankbar, dass ich als LIFE-Fellow die Möglichkeit hatte, an einem intensiven internationalen Austausch über die Vielfalt sozial- und naturwissenschaftlicher Forschung teilzunehmen.

Mein herzlicher Dank gilt außerdem den Mitarbeiterinnen und Mitarbeitern des Instituts zur Qualitätsentwicklung im Bildungswesen für ihre Unterstützung, ihre hilfreichen Rückmeldungen und die angenehme Arbeitsatmosphäre. Insbesondere möchte ich mich bei den Mitgliedern der Agraphiegruppe bedanken, die mich auch in schwierigen Phasen motiviert haben, an der Promotion weiterzuarbeiten. Ein herzlicher Dank geht auch an Prof. Dr. Ulrich Schroeders für die Gründung dieser Gruppe und für seine zahlreichen Ratschläge dazu, wie dem Arbeitstag mehr Stunden für wissenschaftliches Arbeiten abzuringen sind. Persönlich bedanken möchte ich mich an dieser Stelle außerdem bei Dr. Katrin Böhme, Maria Engelbert, Dr. Martin Hecht, Dr. Malte Jansen, Dr. Sebastian Kempert, Ricarda Klein, Aleksander Kocaj, Pauline Kohrt, Dr. Poldi Kuhl, Anna Lenski, Alexandra Marx, Claudia Neuendorf, Prof. Dr. Hans Anand Pant, Andrea Prater, Dr. Heino Reimers, Prof. Dr. Dirk Richter, Dr. Camilla Rjosk, Dr. Alexander Robitzsch, Karoline Sachse, Dr. Stefan Schipolowski, Kristin Schotte, Tatjana Taraszow und Dr. Sebastian Weirich.

Zusammenfassung

Die Leistungsheterogenität in der deutschen Schülerschaft ist ein zentrales Ergebnis nationaler und internationaler Schulleistungsstudien. Insbesondere für Schülerinnen und Schüler mit nicht-deutscher Familiensprache konnten bereits in der Grundschule substanzielle Disparitäten in den Fächern Deutsch und Mathematik festgestellt werden. Diese Disparitäten führten zu der Frage, ob die verwendeten Testverfahren möglicherweise zu hohe sprachliche Hürden für Schülerinnen und Schüler mit nicht-deutscher Familiensprache aufweisen und daher nicht ausreichend in der Lage sind, die Leistungsfähigkeit dieser Gruppe valide zu erfassen. Im Rahmen dieses Dissertationsprojekts wurde geprüft, inwiefern die sprachliche Komplexität von Mathematikaufgaben einen benachteiligenden Einflussfaktor auf die Erfassung der Mathematikleistung von Schülerinnen und Schülern mit nicht-deutscher Familiensprache darstellt. Hierfür wurde zunächst der Frage nachgegangen, in wie weit die in nationalen Schulleistungsstudien verwendeten Mathematikaufgaben in der Grundschule für Schülerinnen und Schüler mit nicht-deutscher Familiensprache differenzielle Validität aufweisen. Daran anschließend wurde untersucht, ob sich die spezifischen Leistungsnachteile von Schülerinnen und Schülern mit nicht-deutscher Familiensprache durch die sprachlichen Merkmale der Aufgaben erklären lassen. Diese Forschungsfragen wurden in vier Teilstudien bearbeitet.

In der ersten Teilstudie (*The Role of Academic-Language Features for Reading Comprehension of Language-Minority Students and Students From Low-SES Families*) wurde zunächst betrachtet, inwiefern die bildungssprachlichen Anforderungen von Lesetestaufgaben mit spezifischen Leistungsnachteilen von Kindern mit niedrigem sozioökonomischem Status (SES) bzw. von Kindern mit nicht-deutscher Familiensprache in Beziehung stehen. Es konnte gezeigt werden, dass die bildungssprachlichen Merkmale der Lesetestaufgaben insgesamt stärker mit den spezifischen Leistungsnachteilen von Kindern mit nicht-deutscher Familiensprache zusammenhängen. Insbesondere lange und komplexe Wörter wirken sich stärker auf die Leseleistung der Schülerinnen und Schüler mit nicht-deutscher Familiensprache als auf monolingual deutschsprachige Schülerinnen und Schüler mit niedrigem SES aus.

In der zweiten Teilstudie (*Second Language Learners' Performance in Mathematics: Disentangling the Effects of Academic Language Features*) wurde das Zusammenspiel bildungssprachlicher Merkmale untersucht. Es konnte gezeigt werden, dass sich die einzelnen bildungssprachlichen Merkmale sowohl spezifisch als auch gemeinsam auf die Leistungsnachteile von Kindern mit nicht-deutscher Familiensprache in Mathematik auswirken. Der größte Anteil der itemspezifischen Leistungsnachteile wurde durch mehrere Merkmale gemeinsam aufgeklärt.

Die dritte Teilstudie (*Linguistic Simplification of Mathematics Items: Effects for Language Minority Students in Germany*) untersuchte experimentell, ob und gegebenenfalls unter welchen Bedingungen eine Reduktion der bildungssprachlichen Anforderungen von Mathematikaufgaben der vierten Klasse zu geringeren Leistungsnachteilen von Schülerinnen und Schülern mit nicht-deutscher Familiensprache führt. In dieser Teilstudie konnte gezeigt werden, dass eine sprachliche Vereinfachung der Aufgaben die Leistungsnachteile von Schülerinnen und Schülern mit mittlerer Lesefähigkeit etwas verringern kann. Insgesamt scheint eine sprachliche Vereinfachung jedoch nicht geeignet, um die Leistungsnachteile von Schülerinnen und Schülern mit nicht-deutscher Familiensprache substanziell zu verringern.

In der vierten Teilstudie (*Effects of Mathematics Items' Language Demands for Language Minority Students: Do They Differ Between Grades?*) wurde untersucht, inwiefern sich die Effekte bildungssprachlicher Merkmale von Mathematikaufgaben auf die Mathematikleistungen von Kindern mit nicht-deutscher Familiensprache zwischen der dritten und der vierten Klassenstufe unterscheiden. In dieser Teilstudie konnte festgestellt werden, dass sich die sprachliche Komplexität der Aufgaben vor allem für jüngere Grundschulkinder unabhängig von ihrer Familiensprache benachteiligend auswirkte.

Abstract

Large-scale assessment studies have repeatedly documented performance disadvantages of language minority students in German elementary schools. The substantial achievement gap has led to concerns regarding the validity of large-scale assessment items for language minority students. It may be the case that these performance differences are, in part, due to high language demands of the test items. These items may selectively disadvantage language minority students in the testing situation. This dissertation project investigated the connection between the academic language demands of mathematics test items and the test performance of monolingual students and language minority students. It was determined whether the test items were differentially valid for language minority students. Moreover, the connection between the differential validity and the linguistic complexity of the test items was tested. The dissertation includes four research articles.

In the first research article (*The Role of Academic-Language Features for Reading Comprehension of Language-Minority Students and Students From Low-SES Families*) we examined whether language minority students are more disadvantaged by the linguistic demands of reading items than German monolingual students with low socio-economic status (SES). The findings indicate that language minority students seemed to struggle especially with long and complex words and with long sentences. Fewer features were found to disadvantage German monolingual students from low SES families and the correlations were generally lower for this group than for language minority students.

In the second research article (*Second Language Learners' Performance in Mathematics: Disentangling the Effects of Academic Language Features*) we examined the interplay of different features of linguistic complexity in their effects on performance disadvantages of language minority students in mathematics. The findings indicate that the largest proportion of item-specific performance disadvantages was explained by confounded combinations of several linguistic features. Moreover, we identified unique effects of descriptive, lexical, and grammatical features.

In the third research article (*Linguistic Simplification of Mathematics Items: Effects for Language Minority Students in Germany*) we experimentally simplified linguistic features of mathematics test items and tested the effects of this simplification for language minority students with varying levels of SES and reading proficiency. Differential effects for language minority students emerged when we took the moderator effects of SES and language proficiency into account, indicating that some language minority students may profit from lin-

guistic simplification in elementary school. However, the overall reduction of performance disadvantages was not substantial.

In the fourth research article (*Effects of Mathematics Items' Language Demands for Language Minority Students: Do They Differ Between Grades?*) we investigated whether mathematics items language demands have differential effects for language minority students over two adjacent grade levels. The findings indicate that the impact of academic language demands seemed to depend on grade level rather than on language minority student status. Regardless of their home language, younger students seemed to struggle more with linguistically complex test items than older students.

Liste der Beiträge

Erster Einzelbeitrag:

Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, *50*(1), 61–82. doi:10.1002/rrq.83

Zweiter Einzelbeitrag:

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24–34. doi:10.1016/j.learninstruc .2013.04.001

Dritter Einzelbeitrag:

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: effects for language minority students in Germany. *European Journal of Psychology of Education, 30*(2), 145-167. doi:10.1007/s10212-014-0233-6

Vierter Einzelbeitrag:

Haag, N., Roppelt, A., & Heppt, B. (2015). Effects of mathematics items' language demands for language minority students: Do they differ between grades? *Learning and Individual Differences, 42,* 70-76. doi: 10.1016/j.lindif.2015.08.010

1 Einleitung

Seit den späten neunziger Jahren haben große Schulleistungsstudien in Deutschland an Bedeutung gewonnen. Mittlerweile gibt es eine Vielzahl dieser Studien wie das *Programme for International Student Assessment* (PISA; Prenzel, Sälzer, Klieme & Köller, 2013), die *Progress in International Reading Literacy Study*[1] (PIRLS; Bos, Tarelli, Bremerich-Vos & Schwippert, 2012), die *Trends in International Mathematics and Science Study* (TIMSS; Bos, Wendt, Köller & Selter, 2012) und die IQB-Ländervergleiche (Pant et al., 2013; Stanat, Pant, Böhme & Richter, 2012). Nationale und internationale Schulleistungsstudien werden regelmäßig bundesweit durchgeführt und ihre Ergebnisse werden in den Medien breit rezipiert. Diese Studien geben wertvolle Hinweise auf die Kompetenzen, die in Deutschland lebende Schülerinnen und Schüler im internationalen und im nationalen Vergleich erreichen und sollen unter anderem relevante Daten für Steuerungsentscheidungen im Bildungssystem bereitstellen.

Ein zentrales Ergebnis aller in Deutschland durchgeführten nationalen und internationalen Schulleistungstests ist die große Leistungsheterogenität in der Schülerschaft. Diese Heterogenität kann zum Teil auf familiäre Merkmale der Schülerinnen und Schüler wie den Zuwanderungshintergrund (Gebhardt, Rauch, Mang, Sälzer & Stanat, 2013; Haag, Böhme & Stanat, 2012; Pöhlmann, Haag & Stanat, 2013; Schwippert, Wendt & Tarelli, 2012; Stanat, Rauch & Segeritz, 2010; Tarelli, Schwippert & Stubbe, 2012) und den sozioökonomischen Status (SES) der Familien (Ehmke & Jude, 2010; Kuhl, Siegle & Lenski, 2012; K. Müller & Ehmke, 2013; Richter, Kuhl & Pant, 2012; Stubbe, Tarelli & Wendt, 2012; Wendt, Stubbe & Schwippert, 2012) zurückgeführt werden. In allen genannten Schulleistungsstudien zeigte sich der stabile Befund, dass Kinder mit Zuwanderungshintergrund in ihren Leistungen hinter ihren Mitschülerinnen und Mitschülern ohne Zuwanderungshintergrund zurückbleiben, wobei sich die Disparitäten in Bezug auf die Lesekompetenz (Stanat et al., 2010) und die mathematische Kompetenz (Gebhardt et al., 2013; Tarelli et al., 2012) über die letzten Jahre etwas verringert haben.

Zur Erklärung dieser Disparitäten wurden vielfach die spezifischen Herausforderungen in den Blick genommen, denen Kinder und Jugendliche aus zugewanderten Familien im Bildungssystem begegnen (für eine Zusammenfassung siehe z. B. Stanat, 2006). Beispielsweise verfügen zugewanderte Familien im Allgemeinen über einen geringeren sozioökonomischen Status, der wiederum mit Bildungserfolg in Zusammenhang steht (Ehmke & Jude, 2010; Kuhl et al., 2012; K. Müller & Ehmke, 2013; Richter et al., 2012; Stubbe et al., 2012; Wendt et

[1] Im deutschen Sprachraum wird diese Studie als *Internationale Grundschul-Lese-Untersuchung* (IGLU) bezeichnet.

al., 2012). Allerdings ist der Effekt des Zuwanderungshintergrunds auf den Bildungserfolg nicht vollständig durch den durchschnittlich niedrigeren SES der Familien zu erklären. Sowohl in den Bereichen Lesen, Mathematik und Naturwissenschaften in der Sekundarstufe (Gebhardt et al., 2013; Pöhlmann et al., 2013; Stanat et al., 2010) als auch in den Bereichen Lesen, Zuhören und Mathematik in der Primarstufe (Haag et al., 2012) bleiben auch nach Kontrolle des SES noch signifikante Leistungsrückstände von Schülerinnen und Schülern mit Zuwanderungshintergrund bestehen.

Ferner weisen Schülerinnen und Schüler mit Zuwanderungshintergrund häufig niedrigere Sprachkenntnisse in der Instruktionssprache auf (Haag et al., 2012; Schwippert et al., 2012; Stanat et al., 2010). Diese geringeren Sprachkenntnisse können sich negativ auf schulische Lernprozesse auswirken und können somit zu kumulativen Lerndefiziten in Sachfächern wie beispielsweise Mathematik führen (Baumert & Schümer, 2001; Brown, 2005; Kölbl, Tiedemann & Billmann-Mahecha, 2006; Stanat, 2006). Aus theoretischer Sicht sollte die Familiensprache ein guter Prädiktor für Sprachkenntnisse in der Instruktionssprache sein, da durch die Verwendung von Deutsch als Familiensprache außerschulische Lerngelegenheiten für den Erwerb sprachlicher Kompetenzen in der Zweitsprache geschaffen werden. Schülerinnen und Schüler mit nicht-deutscher Familiensprache[2] sind dementsprechend stärker auf institutionell geschaffene Lerngelegenheiten zum Erwerb der deutschen Sprache angewiesen (vgl. Stanat, 2006). Allerdings werden in der Schule häufig nicht genügend strukturierte Lerngelegenheiten angeboten, um die sprachlichen Rückstände von Kindern mit nicht-deutscher Familiensprache insbesondere in Bezug auf die Verwendung schulspezifischer sprachlicher Strukturen auszugleichen (vgl. Gogolin, 2009, siehe auch Abschnitt 2.3.2). Effekte des Zuwanderungshintergrunds auf Schulleistungen fielen dementsprechend deutlich geringer aus, wenn für die Verwendung der Instruktionssprache Deutsch als Familiensprache kontrolliert wurde (Haag et al., 2012; Stanat et al., 2010). Trotz der Kontrolle der Familiensprache blieben signifikante Leistungsnachteile für einige Zuwanderungsgruppen bestehen, was darauf hindeuten könnte, dass die mit einer nicht-deutschen Familiensprache zusammenhängenden Leistungsnachteile nicht vollständig durch die in der Schule angebotenen Lerngelegenheiten kompensiert werden können.

Der starke Zusammenhang der Familiensprache mit dem Abschneiden von Schülerinnen und Schülern mit Zuwanderungshintergrund in Schulleistungstests für Sachfächer wie Mathematik und Naturwissenschaften führte auch zu der Frage, ob die verwendeten Testin-

[2] Der Begriff Schülerinnen und Schüler mit nicht-deutscher Familiensprache bezeichnet in der vorliegenden Arbeit nicht ausschließlich monolingual deutschsprachige Schülerinnen und Schüler. Somit werden auch Schülerinnen und Schüler zu dieser Gruppe gezählt, die in ihren Familien sowohl deutsch als auch eine andere Sprache sprechen.

strumente für Schülerinnen und Schüler, deren Familiensprache nicht der Testsprache entspricht, vergleichbare Messgüte aufweisen wie für monolingual aufwachsende Schülerinnen und Schüler. Vor allem in den USA hat die Frage nach der Vergleichbarkeit von Leistungsmessungen für heterogene Gruppen im Zuge des No Child Left Behind (NCLB)-Gesetzes für eine intensive und immer noch andauernde Diskussion gesorgt, die sich beispielsweise in entsprechenden Themenheften der Zeitschriften *Educational Research and Evaluation* (2013), *Applied Measurement in Education* (2014) oder *Review of Research in Education* (2015) widerspiegelt. Bei dieser Diskussion wurden—neben Schülerinnen und Schülern mit Behinderungen—vor allem English Language Learners (ELLs) in den Fokus genommen. Unter ELLs werden Schülerinnen und Schüler mit Zuwanderungshintergrund verstanden, die aufgrund geringer Sprachkenntnisse der englischen Sprache nicht ohne weiteres in der Lage sind, dem normalen Unterricht zu folgen. Das NCLB-Gesetz sah vor, dass auch die von ELLs erzielten Testergebnisse in schulbezogenen Leistungstests herangezogen werden sollen, um die Arbeit von Schulen zu evaluieren. Verschiedene Studien konnten zeigen, dass die dafür verwendeten Tests—die meist ausschließlich an monolingual englischsprachigen Schülerinnen und Schülern erprobt und normiert wurden—für ELLs geringere Reliabilitäten aufweisen (Abedi, 2002; Abedi, Lord & Plummer, 1997; Young et al., 2008). Insbesondere die Validität der Testergebnisse in nicht primär sprachlichen Domänen wie Mathematik wurde angezweifelt, da die Testaufgaben möglicherweise zu hohe sprachliche Hürden aufweisen (z. B. Abedi, 2002; Abedi et al., 1997). Bislang konnte jedoch noch nicht eindeutig festgestellt werden, unter welchen Bedingungen die sprachlichen Anforderungen von Testaufgaben zu einer Benachteiligung von Zweitsprachlernenden[3] führen (für eine aktuelle Zusammenfassung der Befunde siehe Lane & Leventhal, 2015, vgl. auch Abschnitt 2.4 der vorliegenden Arbeit). Im Zusammenhang mit der Frage, ob die sprachlichen Anforderungen von Testaufgaben sich insbesondere für Zweitsprachlernende auswirken, wurde auch diskutiert, ob die Testaufgaben für diese Gruppe von Schülerinnen und Schülern sprachlich vereinfacht werden sollten. Durch eine solche Testanpassung könnten auf sprachliche Anforderungen der Testaufgaben zurückzuführende Leistungsnachteile ausgeglichen werden (z. B. Abedi, Hofstetter & Lord, 2004; Abedi & Lord, 2001; Kieffer, Rivera & Francis, 2012; Pennock-Roman & Rivera, 2011).

Viele der internationalen Untersuchungen der Validität von Testverfahren für ELLs beziehen sich auf den Bereich der Sekundarstufe, insbesondere auf die Fächer Mathematik und Naturwissenschaften (zusammenfassend siehe Lane & Leventhal, 2015). Im deutschsprachigen Raum werden hingegen zunehmend die sprachlichen Anforderungen der Grund-

[3] Der Begriff Zweitsprachlernende bezeichnet in der vorliegenden Arbeit Schülerinnen und Schüler, die in ihren Familien nicht ausschließlich die Instruktionssprache des jeweiligen Landes verwenden.

schulzeit in den Blick genommen (Heppt, Dragon, Berendes, Stanat & Weinert, 2012; Heppt, Stanat, Dragon, Berendes & Weinert, 2014; Verboom, 2008). In diesen Studien konnte gezeigt werden, dass Schülerinnen und Schüler bereits in der Grundschule auf komplexe sprachliche Strukturen treffen und diese verstehen müssen (Berendes, Dragon, Weinert, Heppt & Stanat, 2013; Snow & Uccelli, 2009; Verboom, 2008). Dabei konnte in einigen Studien festgestellt werden, dass Schülerinnen und Schüler mit nicht-deutscher Familiensprache häufiger Probleme haben, diese Strukturen zu verstehen als monolingual deutschsprachige Schülerinnen und Schüler (Berendes et al., 2013; Rösch & Paetsch, 2011). Daher besteht Grund zu der Vermutung, dass Schülerinnen und Schüler mit nicht-deutscher Familiensprache schon im Grundschulalter durch die sprachlichen Anforderungen von Testaufgaben benachteiligt sein könnten. Diese Benachteiligung könnte möglicherweise zur Erklärung der für diese Schülergruppe festgestellten Disparitäten beitragen.

1.1 Anliegen der Arbeit

Die vorliegende Dissertation geht der Frage nach, inwieweit die sprachlichen Anforderungen der Aufgaben deutscher nationaler Schulleistungsstudien in Mathematik und Lesen in der Grundschule besondere Hürden für Kinder mit nicht-deutscher Familiensprache darstellen, wobei der Fokus der vorliegenden Arbeit auf dem Bereich Mathematik liegt. Dieser Fokus wurde gewählt, um die Studie an den Forschungsstand internationaler Studien für die Sekundarstufe anknüpfen zu können. Diese Studien beschäftigen sich vor allem mit den sprachlichen Anforderungen von Aufgaben in den Bereichen Mathematik und Naturwissenschaften, von denen in der Grundschule dem Bereich Mathematik in nationalen Schulleistungsstudien wie dem IQB-Ländervergleich (Stanat et al., 2012) höhere Aufmerksamkeit zukommt. Die vorliegende Arbeit befasst sich zum einen mit dem Ausmaß differenzieller Validität von in nationalen Schulleistungsstudien verwendeten Mathematiktests für die Grundschule für Kinder mit nicht-deutscher Familiensprache. Zum anderen untersucht sie korrelativ und experimentell, inwiefern diese differenzielle Validität durch sprachliche Charakteristika der Aufgaben zu erklären ist. Durch die Analysen soll geprüft werden, inwiefern ein schlechteres Abschneiden von Schülerinnen und Schülern mit nicht-deutscher Familiensprache auf eine differenzielle Validität der verwendeten Testitems zurückgeführt werden kann und ob differenziell valide Items Kinder mit nicht-deutscher Familiensprache vor besondere sprachliche Herausforderungen stellen.

1.2 Gliederung der Arbeit

Zunächst werden kurz einige Modelle des Lösens von mathematischen Textaufgaben vorgestellt (Abschnitt 2.1). In diesem Rahmen wird auch diskutiert, inwiefern die sprachlichen

Anforderungen der Aufgaben beim Lösen von Textaufgaben relevant sind. Im darauffolgenden Abschnitt wird die Beziehung zwischen Validität und Fairness von Leistungsmessungen erläutert (Abschnitt 2.2) und es werden Ansätze zur Erfassung von Testfairness auf der Ebene des Gesamttests und auf der Ebene einzelner Items dargestellt (Abschnitt 2.2.1). In Abschnitt 2.3 wird das Konstrukt der Bildungssprache als Grundlage zur Bestimmung sprachlicher Anforderungen von Testaufgaben vorgestellt. Hierbei werden zunächst einige Annahmen über die Unterschiede zwischen alltäglicher und schulischer Sprache dargestellt, wobei ein Fokus auf die Arbeiten von J. Cummins (1979, 2008) und Schleppegrell (2001, 2004, 2012a) zur Bildungssprache gelegt wird (Abschnitt 2.3.1). Daran anschließend werden familiäre und institutionelle Lerngelegenheiten zum Erwerb bildungssprachlicher Fähigkeiten beschrieben (Abschnitt 2.3.2) und die Messung bildungssprachlicher Merkmale von Testaufgaben erläutert (Abschnitt 2.3.3). Darauf aufbauend werden in Abschnitt 2.4 Befunde zum Einfluss bildungssprachlicher Merkmale von Testaufgaben auf die Testleistungen von Zweitsprachlernenden vorgestellt und diskutiert. Ausgehend von den im theoretischen Teil vorgestellten Forschungsergebnissen werden in Abschnitt 3 die Fragestellungen der vorliegenden Arbeit entwickelt und die vier einzelnen Teilstudien vorgestellt. In Abschnitt 4.1 werden zunächst die zentralen empirischen Befunde der Teilstudien kurz zusammengefasst und in den Forschungsstand eingeordnet. Abschnitt 4.2 thematisiert die Grenzen der Arbeit. Anschließend werden die praktischen (Abschnitt 4.3) und forschungsbezogenen Implikationen der Arbeit diskutiert (Abschnitt 4.4). Zum Abschluss werden einige Forschungsfragen dargestellt, die sich an die vorliegende Arbeit anschließen (Abschnitte 4.4.1 bis 4.4.3). Die vier einzelnen Teilstudien sind im Anhang der Arbeit zu finden.

2 Theoretischer Rahmen der Arbeit

2.1 Modelle des Lösens von Textaufgaben

Die Kognitionspsychologie beschäftigt sich schon seit den 1970er Jahren mit der Frage, wie an alltäglichen Situationen orientierte Textaufgaben gelöst werden. Im Gegensatz zu reinen Arithmetikaufgaben müssen sich Schülerinnen und Schüler bei Textaufgaben zunächst mit dem sprachlich formulierten Aufgabentext auseinandersetzen. Die Rolle der Sprache für das Lösen von Mathematikaufgaben wurde besonders ausführlich für den Bereich des mathematischen Problemlösens in der frühen Grundschulzeit beschrieben (D. D. Cummins, 1991; D. D. Cummins, Kintsch, Reusser & Weimer, 1988; Kintsch, 1986; Kintsch & Greeno, 1985; Reusser, 1997; Stern, 1992). Die kognitiven Modelle der Mathematisierung wurden auf der Grundlage eines Sets an einfachen Additions- und Subtraktionsaufgaben, den sogenannten Standardproblemen (z. B. Reusser, 1997) entwickelt. Die Modelle sollen Schwierigkeitsunterschiede zwischen den Aufgaben erklären und basieren auf der Auswertung der Lösungsversuche von Lernenden zu Beginn der Grundschulzeit. Zur Prüfung der Modelle wurden häufig Computersimulationen eingesetzt, welche die Fehler von Lernenden reproduzieren sollten. Daher sind vor allem die frühen kognitiven Modelle stark formalisiert.

Frühe kognitive Modelle fokussieren auf die logisch-mathematische Struktur der Aufgaben und gehen davon aus, dass Aufgabentexte anhand bestimmter Schlüsselwortstrategien direkt in die zu lösenden mathematischen Terme übertragen werden (*direct translation strategy*, vgl. Mayer & Hegarty, 1996). Nach diesem Modell übersetzen Schülerinnen und Schüler Schlüsselwörter wie beispielsweise „die Anzahl von Kunden“ oder „das Doppelte“ direkt in gegebene bzw. gesuchte Mengen oder mathematische Operationen wie beispielsweise „Variable x (Menge für die Anzahl)“ oder „mit zwei malnehmen“ und ignorieren die im Text enthaltene semantische Information weitgehend. Diese nur die linguistischen Oberflächenmerkmale berücksichtigende Strategie ist allerdings wenig erfolgsversprechend, wenn die im Aufgabentext genannten Schlüsselwörter nicht der Strategie zum Lösen des Problems entsprechen (Hegarty, Mayer & Monk, 1995; Mayer & Hegarty, 1996). Beispielsweise lösten Collegestudenten die Aufgabe „Bei Lucky kostet ein Stück Butter 65 Cent. Das sind 2 Cent weniger pro Stück als bei Vons. Wie viel musst du bei Vons bezahlen, wenn du vier Stück Butter kaufst?“ schlechter als die mathematisch äquivalente Aufgabe „Bei Lucky kostet ein Stück Butter 65 Cent. Bei Vons kostet ein Stück Butter 2 Cent mehr als bei Lucky. Wie viel musst du bei Vons bezahlen, wenn du vier Stück Butter kaufst?“ (Mayer & Hegarty, 1996). Dies wird darauf zurückgeführt, dass die Schlüsselwörter „weniger als“ in der ersten Formulierung der Aufgabe nahelegt, dass 2 Cent von 65 Cent abgezogen werden müssen, wohinge-

gen in der zweiten Formulierung durch die Schlüsselwörter „mehr als“ deutlich wird, dass die 2 Cent zu den 65 Cent hinzugezählt werden müssen. Die Lösungsstrategie „Addieren“, die zur richtigen Lösung des Problems führt, entspricht in der ersten Formulierung des Problems also nicht den im Aufgabentext genannten Schlüsselwörtern.

Demgegenüber gehen die eher sprachverstehensorientierten Modelle davon aus, dass zunächst eine sogenannte *Textbasis* erstellt wird (vgl. Abbildung 1). Dabei handelt es sich um eine Repräsentation des Textes, welche die semantischen Beziehungen der einzelnen Textelemente zueinander abbildet (Kintsch, 1986; Kintsch & Greeno, 1985). Diese logisch schematisierten Repräsentationen werden als *Propositionen* bezeichnet. Mögliche Propositionen sind beispielsweise Aussagen über Objekte (z. B. wie viele Objekte eine Objektmenge umfasst) und über Beziehungen zwischen Objekten. Um diese semantische Repräsentation des Textes zu konstruieren, müssen Schülerinnen und Schüler die im Text gegebenen Sätze in Propositionen überführen. Auf Basis dieser Repräsentation werden dann die zur Lösung des Problems benötigten mathematischen Schemata ausgewählt und in einem *mathematischen Problemmodell* miteinander in Beziehung gebracht (vgl. Abbildung 1). Hierbei kann es sich um Schemata zur Repräsentation von Mengen und Beziehungen von Mengen zueinander handeln (z. B. Teil-Ganzes-Schema) oder um Schemata zur Repräsentation von Zähloperationen und arithmetischen Operationen (z. B. Objekte von einer Menge zu einer anderen hinzufügen oder von einer anderen Menge entfernen) (vgl. Kintsch & Greeno, 1985).

Die Schwierigkeit einer Aufgabe kann somit auf zwei Faktoren zurückzuführen sein. Der erste Faktor umfasst die zur Aufgabenstellung passenden Schemata. Schülerinnen und Schüler müssen diese Schemata zum einen bereits erworben haben, um sie bei einer konkreten Aufgabe anwenden zu können. Darüber hinaus müssen die Schülerinnen und Schüler das zur Aufgabe passende Schema aus ihrem Repertoire auswählen und korrekt anwenden (vgl. Stern, 1992). Der zweite Faktor bezieht sich darauf, wie stark die Aufgabe das Arbeitsgedächtnis der Schülerinnen und Schüler fordert. Dieser Faktor wird durch die Anzahl von Propositionen bestimmt, die zunächst keinem Schema zugeordnet werden können, sondern im Arbeitsgedächtnis gehalten werden müssen (Kintsch & Greeno, 1985). Diese zweite durch die Konstruktion der Aufgabe bedingte Schwierigkeit wird auch als *extraneous cognitive load* bezeichnet (Paas, Renkl & Sweller, 2003).

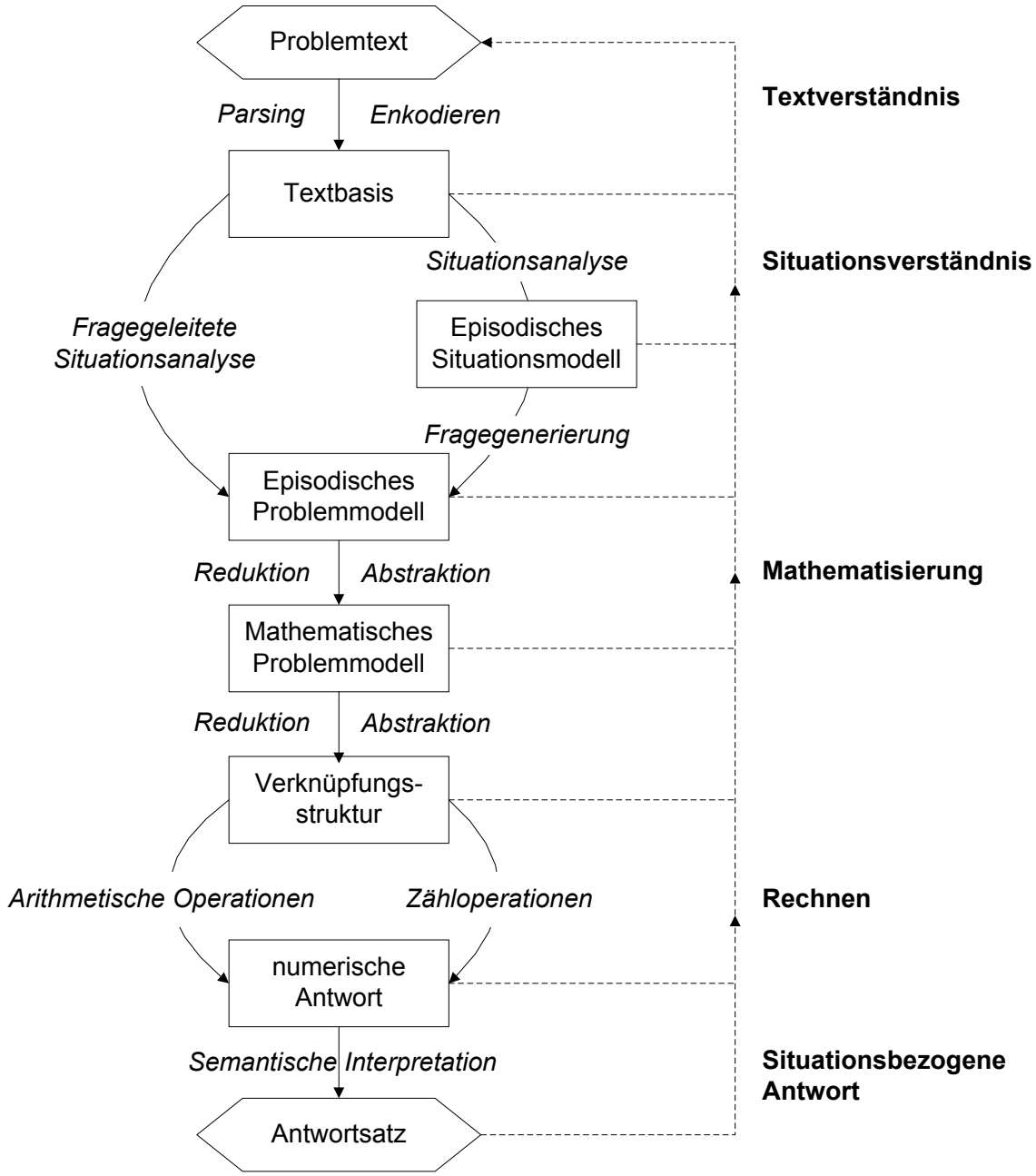

Abbildung 1: Verstehensebenen bzw. Stufen der Mathematisierung von Textaufgaben (angelehnt an Reusser, 1997, S. 151)

Ein idealtypisches Modell des Prozesses, der sich zwischen dem Lesen einer Aufgabe und dem Lösen dieser Aufgabe entfaltet, wurde von Reusser (1997) vorgestellt (Abbildung 1). Dieses Modell integriert die bereits beschriebenen früheren Modelle, die nur einen Teil der Schritte umfassen. Nach Reusser (1997) wird beim Lesen einer Textaufgabe, wie in früheren Modellen beschrieben, zunächst die Oberflächenstruktur in eine *Textbasis* überführt, die die Beziehungen zwischen den Objekten als Propositonen abbildet (Kintsch, 1986; Kintsch & Greeno, 1985). Auf Grundlage dieser Textbasis wird ein handlungsnahes *episodisches Situa-*

tionsmodell der Textaufgabe gebildet, welches unter Zuhilfenahme der Fragestellung in ein abstrakteres *episodisches Problemmodell* überführt wird. Auf dieser Ebene werden die konkreten Handlungen des Aufgabenkontexts repräsentiert. Die episodischen Modelle bilden eine zusätzliche, eher alltagsnahe Verstehensebene zwischen der Textbasis als formale Repräsentation des Textes und dem mathematischen Problemmodell als formale Repräsentation der mathematischen Beziehungen der Variablen. Aus dem episodischen Problemmodell wird sodann ein *mathematisches Problemmodell* abgeleitet. Dieses formalisiert die Zusammenhänge relevanter Variablen mit mathematischen Mitteln und übersetzt die Realität in die Sprache der Mathematik. Durch weitere Reduktion wird eine *mathematische Verknüpfungsstruktur* erstellt, welche die relevanten Operationen zur Lösung der Aufgabe beinhaltet. Werden diese Operationen durchgeführt, ergibt sich eine *numerische Antwort*, die im letzten Schritt der *semantischen Interpretation* wieder auf den ursprünglichen Aufgabenkontext zurückbezogen werden muss.

Das Erstellen von handlungsnahen Situationsmodellen wirkt sich positiv auf das Lösen von Mathematikaufgaben aus (Boonen, van Wesel, Jolles & van der Schoot, 2014; Pape, 2004). Bei bekanntem Problemkontext können einzelne Schritte des Mathematisierungsprozesses übersprungen werden (vgl. Kintsch, 1986; Kintsch & Greeno, 1985). Beispielsweise konzentrieren sich Schüler mit hohen Leistungen in Mathematik beim Verarbeiten von Aufgabentexten eher auf Informationen, die für die Erstellung von Problemmodellen relevant sind, als auf solche, die für die Erstellung von Situationsmodellen relevant sind (Moreau & Coquin-Viennot, 2003). Daher wird vermutet, dass Situationsmodelle vor allem dann relevant sind, wenn entweder die entsprechenden mathematischen Schemata von den Schülerinnen und Schülern noch nicht routiniert beherrscht werden oder es sich um neue Problemtypen handelt (D. D. Cummins, 1991; vgl. auch Stern, 1992; Tolar et al., 2012).

Sprachliche Fähigkeiten sind für das Lösen von Mathematikaufgaben vor allem bei der Übersetzung der Aufgabenstellung von der linguistischen Oberflächenstruktur in ein episodisches Situationsmodell und bei der Integration der gegebenen Informationen in dieses Modell von Bedeutung (vgl. D. D. Cummins et al., 1988; Wu & Adams, 2006). Entsprechend zeigte sich in verschiedenen Studien ein Zusammenhang zwischen sprachlichen Fähigkeiten und der Erstellung mentaler Modelle der Problemsituation (Lynn S. Fuchs et al., 2008; Lee, Ng & Ng, 2009; Pape, 2004). In Untersuchungen mit Neuntklässlerinnen und Neuntklässlern zeigten Leiss, Schukajlow, Blum, Messner und Pekrun (2010), dass nicht die allgemeine Lesekompetenz, sondern vor allem die mit einem mathematischen Lesetest gemessene fachspezifische Lesekompetenz die Lösung von mathematischen Modellierungsaufgaben beeinflusst. Der mathematische Lesetest bestand aus einer Mathematikaufgabe, die neben den zur

Lösung der Aufgabe notwendigen Informationen weitere Informationen umfasste. Die Schülerinnen und Schüler sollten der Aufgabe diejenigen Informationen entnehmen, die zur Lösung notwendig sind, mussten die Aufgabe selbst jedoch nicht lösen.

Besonders jüngere Grundschülerinnen und Grundschüler scheiterten im Aufgabenlöseprozess an Fehlinterpretationen einzelner Wörter. Interpretieren beispielsweise Erstklässler in der Aufgabe „Mary and John have 5 marbles altogether. Mary has 3 marbles. How many marbles does John have?" das Wort „altogether" als „each", können selbst Kinder, die das zur Lösung dieses Problems notwendige Teil-Ganzes-Schema besitzen, die entsprechenden Aufgaben schlechter lösen, da sie Probleme haben, dieses Schema der Aufgabe zuzuordnen (D. D. Cummins, 1991, S. 262).

2.2 Validität und Fairness von Leistungsmessungen

Validität gilt—neben Objektivität und Reliabilität—als ein zentrales Gütekriterium von Testverfahren. Unter Validität wird häufig verstanden, ob der Test misst, was er messen soll (vgl. Hartig, Frey & Jude, 2012). Zur Beantwortung dieser Frage können verschiedene Ansätze verfolgt werden. In der frühen Validitätsforschung wurden daher verschiedene Arten der Validität unterschieden (für einen historischen Überblick siehe Lissitz & Samuelsen, 2007). Zu Beginn der Validitätsforschung in den 1930er Jahren wurden weitgehend atheoretische Zusammenhänge zwischen Tests und interessierenden Kriterien in der Außenwelt (z. B. Berufserfolg) untersucht, um herauszufinden, wie gut diese Kriterien mit Hilfe des Testergebnisses vorhergesagt werden können. Diese Art der Validität wird als *Kriteriumsvalidität* bezeichnet. Eine hohe Kriteriumsvalidität liegt vor, wenn die Testergebnisse eine präzise Vorhersage relevanter Kriterien ermöglichen. In späteren Arbeiten nehmen den durch den Test zu erfassenden theoretischen Konstrukten einen größeren Raum ein. Hierbei werden wiederum verschiedene Validitätsarten unterschieden. Die *Inhaltsvalidität* gibt an, ob der Test alle wesentlichen theoretisch angenommenen Bereiche des Konstrukts abdeckt und wird meist über Expertenurteile erfasst. Theoriegeleitet konstruierte Tests weisen im Allgemeinen eine hohe Inhaltsvalidität auf. Die *Konstruktvalidität* bezeichnet den Grad der Einbettung des zu testenden Konstrukts in ein nomologisches Netzwerk ähnlicher und unterschiedlicher Konstrukte und wird meist über die Unterschiede zwischen Korrelationen des Tests mit bestehenden Testverfahren für diese ähnlichen und unterschiedlichen Konstrukte erfasst. Eine hohe Konstruktvalidität wäre dann gegeben, wenn der Test hoch mit ähnlichen Konstrukten, aber nicht mit unterschiedlichen Konstrukten korreliert (vgl. Hartig et al., 2012).

Aktuelle Validitätskonzepte wurden maßgeblich von der Arbeit von Messick (1989) beeinflusst. Messick schlägt einen integrierten Validitätsbegriff vor, in dem Validität als eine umfassende theoretische und empirische Beurteilung der Bedeutung, Relevanz und Nützlichkeit von Testergebnissen *für konkrete Zwecke* verstanden wird (Messick, 1989, S. 13). Somit ist Validität für Messick keine feste Eigenschaft eines bestimmten Tests, sondern hängt maßgeblich davon ab, für welchen Zweck und unter welchen Umständen der Test eingesetzt werden soll. Die Validität eines Tests muss empirisch und theoretisch für jedes intendierte Einsatzgebiet nachgewiesen werden. Dazu sollten folgende Kriterien geprüft werden: (1) Interpretation/Bedeutung der Testergebnisse, (2) Relevanz/Nützlichkeit der Testergebnisse für den konkreten Anwendungsfall, (3) mit Testergebnissen verbundene Werturteile und Handlungsempfehlungen, (4) intendierte und unintendierte Konsequenzen der Testung. Insbesondere die Konsequenzen der Testung—deren Wichtigkeit Messick betont—werden in traditionellen Konzeptualisierungen von Validität nicht berücksichtigt. Die *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999, 2014) arbeiten diese Kriterien der Validitätsevidenz weiter aus, um sie für die Durchführung von Validierungsstudien besser nutzbar zu machen.

Für die vorliegende Arbeit ist vor allem das erste Kriterium—die Interpretation der Testergebnisse—von zentraler Bedeutung. Dieses Kriterium bezieht sich darauf, dass Unterschiede in den Testergebnissen zwischen Personen ausschließlich auf Unterschiede in der Ausprägung des durch den Tests gemessenen Konstrukts zurückzuführen sein sollen. Die Validität der Testinterpretation ist gefährdet, wenn entweder der Test nicht alle Facetten des Konstrukts erfasst (*construct underrepresentation*) oder die Unterschiede in den Testergebnissen zwischen Personen nicht ausschließlich auf das zu messende Konstrukt, sondern auch auf die Ausprägung anderer Merkmale zurückzuführen sind, die der Test nicht erfassen soll (*construct-irrelevant variance*, Messick, 1989). Eine Beeinflussung der Testwerte durch konstruktirrelevante Varianz könnte beispielsweise vorliegen, wenn ein Schulleistungstest computeradministriert durchgeführt wird und die Schülerinnen und Schüler sich darin unterscheiden, wie gut sie mit dem Computer umgehen können. Die Testwerte wären dann nicht nur vom fachlichen Wissen der Schülerinnen und Schüler beeinflusst—dem Konstrukt, das der Test erfassen soll—, sondern auch von der konstruktirrelevanten Fähigkeit, die Aufgaben am Computer zu bearbeiten. Schülerinnen und Schüler, die bei der Bedienung des Programms Probleme haben, wären durch diese Testadministration benachteiligt und ihre Testwerte würden nicht ihre „wahre“ fachliche Fähigkeit widerspiegeln. Daher könnten die Testwerte in einem solchen Fall nicht als valider Indikator für die fachliche Fähigkeit interpretiert werden.

Während sich die Validitätsforschung vorrangig mit der Gültigkeit und Aussagekraft von Testwerten für die intendierte Testpopulation insgesamt beschäftigt, liegt der Fokus der Fairnessforschung auf der Vergleichbarkeit von Messinstrumenten, Testdurchführung und Testinterpretation für *verschiedene* Gruppen innerhalb der intendierten Testpopulation (Cole & Zieky, 2001). Ähnlich wie Validität ist Fairness nicht ausschließlich eine Eigenschaft des Tests, sondern hängt auch von der Testverwendung in einer konkreten Situation ab.

Bei einer Fairnessprüfung wird für jeden Schritt einer Testvalidierung geprüft, ob dieser Schritt für die einzelnen Gruppen innerhalb der intendierten Testpopulation zu vergleichbaren Aussagen führt. Ein Test wäre somit fair, wenn die Testwerte, die Interpretation dieser Werte, die daraus gezogenen Schlüsse und die Konsequenzen dieser Schlüsse für die einzelnen Gruppen innerhalb einer Testpopulation vergleichbar sind. Differenzielle Aussagen über Testteilnehmer verschiedener Gruppen werden als *differenzielle Validität* des Tests bezeichnet (vgl. Cole & Zieky, 2001; Xi, 2010). Fairness kann also als die Abwesenheit differenzieller Validität für die betrachteten Gruppen bezeichnet werden.

Die Frage nach der Fairness von Testinstrumenten stellt sich vor allem bei Tests, die in heterogenen Stichproben eingesetzt werden sollen. Heterogene Stichproben umfassen Testteilnehmer, die sich hinsichtlich verschiedener Merkmale unterscheiden. In Schulleistungstests kann die Schülerschaft als heterogene Stichprobe angesehen werden, da sich die Schülerinnen und Schüler in ihren Hintergrundmerkmalen wie ihrem Zuwanderungshintergrund, dem SES ihrer Familien oder der besuchten Schulform unterscheiden (Schwabe & Gebauer, 2013). Diese Merkmale können Quellen konstruktirrelevanter Varianz sein, die dazu führen, dass die Testwerte für einzelne Gruppen nicht mehr als valide Indikatoren ihrer Fähigkeit interpretiert werden können. Daher sollte im Rahmen einer Fairnessprüfung festgestellt werden, inwieweit Unterschiede zwischen den Testleistungen einzelner Gruppen innerhalb einer heterogenen Population ausschließlich auf die Unterschiede in der Ausprägung des zu messenden Konstrukts zurückzuführen sind. In der Betrachtung der Fairness von Schulleistungstests ist vor allem wichtig, dass Unterschiede zwischen interessierenden gesellschaftlichen Gruppen sowie die Veränderungen dieser Unterschiede über die Zeit valide abgebildet werden und nicht auf eventuelle Unfairness des Testinstruments zurückgeführt werden können (vgl. Zumbo, 2007). Für die Fairnessprüfung werden meist die Unterschiede zwischen zwei Gruppen betrachtet, die anhand eines interessierenden manifesten Merkmals wie beispielsweise Geschlecht oder Zuwanderungshintergrund gebildet wurden. Bei entsprechend großen Stichproben kann jedoch auch eine Kombination von manifesten oder latenten Merkmalen zur Gruppenbildung herangezogen werden, um zu möglichst homogenen Gruppen zu gelangen (*DIF dissection*, vgl. Ercikan & Oliveri, 2013).

2.2.1 Erfassung von Fairness

Bei der Betrachtung von Testfairness können zwei Ebenen unterschieden werden: die Gleichbehandlung bezüglich des Testmaterials, der Testauswertung und der Testsituation (*prozedurale Ebene der Testfairness*) und die gleichwertige Interpretation von Testergebnissen und daraus folgenden Konsequenzen (*interpretationsbezogene Ebene der Testfairness*) (Kane, 2010; Schwabe & Gebauer, 2013). Die prozedurale Ebene der Testfairness kann in der Regel durch eine standardisierte Testadministration und Testauswertung sichergestellt werden und muss daher meist nicht empirisch geprüft werden. Die Sicherstellung der interpretationsbezogenen Ebene der Testfairness gestaltet sich hingegen schwieriger, da sich mögliche Effekte der Testung erst nach dem Einsatz herausstellen (Kane, 2010). Die Prüfung der interpretationsbezogenen Testfairness ist daher häufig Gegenstand empirischer Untersuchungen, wobei es verschiedene Möglichkeiten gibt, die Fairness eines Tests zu prüfen. Diese unterscheiden sich zum einen darin, ob die Fairness des Tests im Hinblick auf Gruppen oder Einzelpersonen untersucht wird und zum anderen darin, ob der Test als Ganzes auf Fairness geprüft wird oder ob Gruppenunterschiede in der Bearbeitung der einzelnen Items untersucht werden (Sireci & Rios, 2013). Da Aussagen über Einzelpersonen in Schulleistungsstudien nicht angestrebt werden, beschränkt sich die Fairnessprüfung in diesem Fall zumeist auf die Unterschiede zwischen Gruppen.

Um die Fairness eines Gesamttests zu prüfen, können drei verschiedene Arten von Analysen durchgeführt werden. Noch vor der Erhebung empirischer Daten kann das Testinstrument von Experten begutachtet werden. Diese Begutachtung bezieht sich zumeist auf mögliche Nachteile einzelner Gruppen in der Bearbeitung des Instruments, die durch institutionelle oder sprachliche Gegebenheiten bedingt werden. Beispielsweise könnte ein Mathematiktest für Schülerinnen und Schüler an Hauptschulen differenziell valide sein, wenn bestimmte Testinhalte zwar in Realschulen und Gymnasien in der getesteten Klassenstufe bereits unterrichtet wurden, in Hauptschulen jedoch (noch) nicht. Dies trifft beispielsweise auf die Kreisberechnung zu, die in Realschulen und Gymnasien üblicherweise in der 8. Klasse bereits im Unterricht behandelt wurde, in Hauptschulen jedoch noch nicht. In diesem Fall würden sich Unterschiede in den Testwerten bei Aufgaben zur Kreisberechnung nicht nur auf Unterschiede in der Mathematikfähigkeit zwischen den Schülerinnen und Schülern zurückführen lassen, sondern würden auch die mangelnden Lerngelegenheiten für Hauptschülerinnen und Hauptschüler widerspiegeln. Diese Art von Fairnessprüfung wird üblicherweise bereits im Konstruktionsprozess von Schulleistungstests durchgeführt (Camilli, 2013).

Empirisch kann geprüft werden, inwiefern die Testwerte für einzelne Gruppen unterschiedlich stark mit relevanten Außenkriterien zusammenhängen. Liegen keine validen Außenkriterien vor oder ist die Messung möglicher Außenkriterien zu aufwändig, kann außerdem geprüft werden, inwiefern sich die Umrechnung der Rohtestwerte auf eine bereits etablierte Berichtsmetrik für die Gruppen voneinander unterscheidet. Eine unterschiedliche Transformationsvorschrift würde darauf hindeuten, dass die im Test erzielte Punktzahl für die einzelnen Gruppen zu unterschiedlichen Werten auf der Berichtsmetrik führt und der Test somit für die einzelnen Gruppen zu unterschiedlichen Schlüssen führen kann (vgl. Lane & Leventhal, 2015). Der Vergleich der Transformationsvorschriften wird im Rahmen sogenannter *Score Equity Analysen* durchgeführt (vgl. Dorans, 2004). Für diese Verfahren ist es notwendig, dass eine Referenzpopulation von Items mit bekannten Itemparametern vorliegt und eine Berichtsmetrik für den Test definiert wurde, deren Fairness nachgewiesen werden konnte. Beides stammt üblicherweise aus einer Normierungsstudie des Tests. Es ist jedoch nicht notwendig, dass exakt derselbe Test wie in der Normierungspopulation durchgeführt wurde, solange die beiden Tests miteinander *verlinkbar* sind (vgl. Dorans, 2004). Zunächst wird der zu verlinkende Test mit Schülerinnen und Schülern der interessierenden Gruppen—dies könnten beispielsweise Schülerinnen und Schüler ohne Zuwanderungshintergrund und Schülerinnen und Schüler mit Zuwanderungshintergrund sein—durchgeführt. Dabei erhält jede Schülerin/jeder Schüler einen Rohpunktwert. Sodann werden die Testergebnisse für die Gruppen getrennt auf die Berichtsmetrik der Referenzpopulation transformiert. Der Test wäre dann unfair, wenn die getrennte Transformation bei gleichem Rohpunktwert einer Schülerin/eines Schülers aus unterschiedlichen Gruppen zu substanziell unterschiedlichen transformierten Skalenwerten führen würde. Huggins und Elbaum (2013) konnten mit diesem Ansatz beispielsweise zeigen, dass eine an die spezifischen Bedürfnisse der Schülerinnen und Schüler angepasste Version eines naturwissenschaftliches Tests für Schülerinnen und Schüler mit Behinderungen oder geringen englischen Sprachkenntnissen die Vergleichbarkeit der Testergebnisse mit Schülerinnen und Schülern ohne diese Beeinträchtigungen für die fünfte Klasse verbessert.

Ein weiterer Ansatz der empirischen Fairnessprüfung konzentriert sich auf die einzelnen Items des Tests. Fairnessprüfungen der Items werden deutlich häufiger durchgeführt als Fairnessprüfungen des Gesamttests (vgl. Young, 2009). Bei der Prüfung von Testfairness auf Itemebene werden die in einem Test gefundenen Gruppenunterschiede zunächst in zwei Teile (*item impact* und *item bias*) zerlegt. *Item impact* bezeichnet Gruppenunterschiede, die sich auf das Konstrukt zurückführen lassen, welches der Test messen soll. Im Gegensatz dazu bezeichnet *item bias* Gruppenunterschiede, die auf konstruktirrelevante Merkmale

zurückgeführt werden können (Ackerman, 1992). Dieser Zerlegung liegt die Idee zugrunde, dass Tests versteckte Mehrdimensionalität aufweisen können (Ackerman, 1992; Camilli, 2006). Versteckte Mehrdimensionalität bezeichnet die Situation, dass ein Test eindimensional konstruiert wurde, aber neben der intendierten Dimension eine oder mehrere konstruktirrelevante Dimensionen miterfasst. Beispielsweise soll ein Mathematiktest die Leistungen von Schülerinnen und Schülern in Mathematik messen. Da die Testaufgaben jedoch sprachlich gestellt werden, misst der Test zu einem gewissen Grad immer auch sprachliche Fähigkeiten. Diese sprachlichen Fähigkeiten werden bei der Modellierung des Mathematiktests als eindimensionales Konstrukt vernachlässigt. Damit ist implizit die Annahme verbunden, dass der Einfluss sprachlicher Fähigkeiten auf die Lösungswahrscheinlichkeit der Items für verschiedene Gruppen, beispielsweise für Kinder mit deutscher Familiensprache und für Kinder mit nicht-deutscher Familiensprache, in gleichem Maße vernachlässigbar ausgeprägt ist. Unterscheidet sich die Auswirkung der sprachlichen Fähigkeiten allerdings zwischen den Gruppen, können Gruppenunterschiede für die Lösungswahrscheinlichkeit der Items gefunden werden, die nicht auf Unterschiede in der Mathematikfähigkeit zurückgeführt werden können. Unterscheiden sich die Items außerdem in ihrer sprachlichen Komplexität, kann sich auch das Ausmaß unterscheiden, in dem die einzelnen Items neben mathematischen auch sprachliche Fähigkeiten erfassen. Diese auf die nicht modellierte konstruktirrelevante Dimension zurückführbaren Gruppenunterschiede in den Lösungshäufigkeiten werden als *item bias* bezeichnet. Liegt *item bias* vor, werden die Ergebnisse des eigentlich eindimensional konstruierten Tests von mindestens einer weiteren, nicht intendierten, Dimension beeinflusst, deren Ausprägung sich zwischen den Gruppen unterscheidet. Daher gefährdet *item bias* im Gegensatz zu *item impact* die Testfairness.

Um die Effekte von *item impact* und *item bias* auf der Ebene der Einzelitems statistisch voneinander zu trennen, werden meist Analysen des differenziellen Itemfunktionierens (*Differential Item Functioning*, DIF) vorgenommen (siehe Camilli, 2006; Osterlind & Everson, 2009; Sireci & Rios, 2013 für einen Überblick von DIF-Methoden). DIF für ein bestimmtes Item liegt vor, wenn sich die gruppenspezifischen Lösungswahrscheinlichkeiten für dieses Item unter Kontrolle der Personenfähigkeit unterscheiden und gibt ein Indiz darauf, dass das Item für die Gruppe mit der geringeren Lösungswahrscheinlichkeit unfair sein könnte. Der für das zu kontrollierende Fähigkeitsmaß festgestellte Gruppenunterschied entspricht hierbei dem *item impact*, wohingegen die nach der Kontrolle des Fähigkeitsmaßes bestehenden Unterschiede in den itemspezifischen Lösungswahrscheinlichkeiten dem *item bias* entsprechen. Der Wahl des Fähigkeitsmaßes, für das kontrolliert wird, kommt hierbei eine hohe Bedeutung zu. Das Fähigkeitsmaß sollte valide und fair für die betrachteten Gruppen

sein. Ein unfaires Fähigkeitsmaß kann dazu führen, dass die Gruppenunterschiede in der Fähigkeit nicht ausreichend genau von den DIF-Werten getrennt werden können. Meist wird der Gesamttest oder ein als für beide Gruppen valide erachteter Teil des Gesamttests als Maß für die Fähigkeit herangezogen (Sireci & Rios, 2013).

DIF-Analysen wurden ursprünglich dazu benutzt, Items mit Bias gegen benachteiligte Gruppen zu identifizieren und aus dem Test auszuschließen, um möglichst faire Aussagen über Gruppenunterschiede auf dem gemessenen Konstrukt zu ermöglichen (für einen historischen Überblick siehe z. B. Zumbo, 2007). Zu beachten ist allerdings, dass differenziell funktionierende Items nur dann aus dem Test ausgeschlossen werden sollten, wenn sich eine inhaltlich plausible Erklärung für den DIF finden lässt (Sireci, Han & Wells, 2008). Zu diesem Zweck können die DIF-Werte statistisch mit konstruktirrelevanten schwierigkeitsgenerierenden Itemmerkmalen in Beziehung gebracht werden. Ferne und Rupp (2007) kritisieren die bisherigen Ergebnisse bezüglich der Ursachen von DIF in Sprachtests allerdings als wenig aufschlussreich: "Despite the laudable attempts to account for causes for DIF, the methodological heterogeneity of studies along with the general lack of explanatory power of predictor variables has made the practical utility of many DIF studies for item revision and development questionable. In other words, it is very difficult to say what conclusions can be *reliably* drawn from 15 years of research on DIF in language testing for the *future* construction of tests that measure a specific ability." (Ferne & Rupp, 2007, S. 145, Hervorhebungen im Original).

Neuere DIF-Analysen fokussieren daher zunehmend darauf, die dem DIF zugrundeliegenden Dimensionen oder kognitiven Prozesse zu identifizieren, zu modellieren und zu verstehen. Der hypothesenprüfende Ansatz der neueren DIF-Forschung lässt sich besonders gut mit mehrebenenanalytischen Methoden untersuchen, insbesondere mit logistischen Regressionsanalysen oder explanatorischen Item Response Modellen (van den Noortgate & De Boeck, 2005). Diese Modelle basieren auf der Item Response-Theorie und bieten die Möglichkeit, Item- und Personenprädiktoren direkt zur Erklärung von Varianz in den Itemantworten heranzuziehen (De Boeck & Wilson, 2004). Explanatorische Item Response Modelle stellen eine Verallgemeinerung von hierarchischen Modellen mit kreuzklassifizierter Struktur dar, in denen die Itemantworten sowohl in Personen als auch in Items genestet sind (Beretvas, Cawthon, Lockhart & Kaye, 2012; De Boeck et al., 2011). In diesen Modellen wird DIF als die Interaktion eines erklärenden Itemmerkmals (z. B. sprachliche Komplexität der Items) und eines Gruppierungsmerkmals der Personen (z. B. Familiensprache) modelliert. Im Vergleich zur nachträglichen Erklärung von gemessenem DIF durch Regressions- oder Varianzanalysen wird in explanatorischen Item Response-Modellen nicht für jedes einzelne

Item ein DIF-Wert errechnet. Stattdessen wird untersucht, inwiefern sich ein Itemmerkmal differenziell auf die Itemantworten einer Gruppe von Personen auswirkt (Beretvas et al., 2012; Isaac & Hochweber, 2011). Dieses Vorgehen führt sowohl zu höherer Teststärke (Beretvas et al., 2012) als auch zu besser interpretierbaren Befunden für die Effekte von Itemmerkmalen (Xie & Wilson, 2008). Daher bietet die Verwendung von explanatorischen Item Response-Modellen eine Möglichkeit, der Kritik von Ferne und Rupp (2007) an bisherigen DIF-Untersuchungen entgegenzutreten und die explanatorische Power von Itemmerkmalen zumindest aus statistischer Sicht zu erhöhen.

2.3 Bildungssprache

Sprachliche Anforderungen von Testaufgaben können mit dem Konzept der Bildungssprache systematisch beschrieben werden. Bildungssprache bezeichnet diejenige Sprache, die im Klassenzimmer oder in anderen akademischen Kontexten verwendet wird, um Wissen zu erwerben und weiterzugeben (Chamot & O'Malley, 1994). Auf linguistischer Ebene unterscheidet sich diese Sprache sowohl in lexikalischen als auch in grammatischen Merkmalen substanziell vom alltäglichen Sprachgebrauch. Beschreibungen von Bildungssprache basieren zumeist auf den linguistischen Merkmalen von Schulbuchtexten oder Unterrichtsgesprächen, wobei dem Wortschatz in der Forschung mehr Aufmerksamkeit gewidmet wurde als grammatischen Merkmalen. Weitere Ansätze beziehen außerdem kognitive und soziokulturell/psychologische Merkmale (Scarcella, 2003; Snow & Uccelli, 2009) sowie metasprachliche Fähigkeiten der Schülerinnen und Schüler, beispielsweise morphologische Bewusstheit (Uccelli, Barr, et al., 2015), mit ein.

Das Konzept der Bildungssprache spielt eine zentrale Rolle bei der Frage, warum Zweitsprachlernende mit bestimmten sprachlichen Merkmalen von Testaufgaben größere Probleme haben könnten als Erstsprachlernende. Dabei wird von zwei Grundannahmen ausgegangen: Zum einen wird postuliert, dass die Beherrschung der Bildungssprache enger mit schulischem Erfolg zusammenhängt als die Beherrschung grundlegender alltagssprachlicher Kompetenzen (Bailey, Butler, Stevens & Lord, 2007; J. Cummins, 1979; Gogolin, 2009; Schleppegrell, 2004). Zum anderen gilt Bildungssprache als besondere Hürde für Kinder mit Zuwanderungshintergrund oder Kinder aus bildungsfernen Familien. Diese Annahme beruht darauf, dass außerschulischen Lerngelegenheiten eine große Rolle zur Entwicklung bildungssprachlicher Fähigkeiten zugeschrieben wird und dass im alltäglichen Umfeld von Kindern mit Zuwanderungshintergrund oder Kindern aus bildungsfernen Familien weniger Lerngelegenheiten für den Erwerb der Bildungssprache bereit gestellt werden (Carhill, Suarez-Orozco & Paez, 2008; Gogolin & Lange, 2011; Uesseler, Runge & Redder, 2013). Außerschulische Lerngelegenheiten spielen für den Erwerb der Bildungssprache deshalb eine

große Rolle, weil Bildungssprache oberflächlich eine große Ähnlichkeit zur Alltagssprache aufweist und daher—mit Ausnahme der Fachsprache—in der Schule selten thematisiert wird (Ehlich, 1999; Pierce & Melena, 2009; Verboom, 2008). Beide Annahmen sind bislang jedoch nicht umfassend empirisch überprüft.

2.3.1 Theoretische Annahmen

Nach gängiger Auffassung dient Bildungssprache dazu, kognitiv komplexe Inhalte in Situationen zu vermitteln, in denen Sprecher und Empfänger keinen geteilten Kontext zur Interpretation des Gesagten haben (z. B. Snow & Uccelli, 2009; Wong Fillmore & Snow, 2002). Dieses Verständnis von Bildungssprache wurde wesentlich von J. Cummins (1979) geprägt, der zwei Bereiche sprachlicher Fähigkeiten unterscheidet. Der erste Bereich—die sogenannten *Basic Interpersonal Communication Skills* (BICS)—umfasst die Kommunikation in alltäglichen, kontextualisierten Interaktionen. Dieser Bereich wird von Zweitsprachlernenden meist relativ schnell erlernt, sodass sie schon nach relativ kurzer Zeit keine Förderung mehr zu benötigen scheinen. Cummins argumentiert jedoch, dass sich die in der Schule verwendete Sprache systematisch von der im Alltag verwendeten Sprache unterscheidet, sodass diese Kinder dennoch Probleme haben können, dem Unterricht zu folgen und angemessen an Interaktionen in Lehr-Lern-Kontexten teilzunehmen. Der zweite Bereich umfasst daher die für sprachliche Interaktionen in Lehr-Lern-Kontexten benötigten sprachlichen Fähigkeiten und wird als *Cognitive/Academic Language Proficiency* (CALP) bezeichnet. Im Gegensatz zu BICS wird CALP eher dazu benötigt, kognitiv komplexe Inhalte in dekontextualisierten Interaktionssituationen zu verstehen und zu vermitteln. In diesen dekontextualisierten Situationen müssen zur thematischen Einordnung des Gesagten bzw. Geschriebenen notwendige Informationen expliziert werden, da nicht, wie in alltäglichen Konversationssituationen, auf para- und nonverbale Signale zurückgegriffen werden kann und eine gemeinsame Wissensbasis von Autor und Leser eines Textes nicht vorausgesetzt werden kann.

Die Unterscheidung von BICS und CALP war zwar wegweisend für die Entwicklung des Konzepts der Bildungssprache, wurde jedoch auch häufig kritisiert (Aukerman, 2007; Faltis, 2013; Gee, 2014; MacSwan, 2000; Schleppegrell, 2004). Die Kritik bezieht sich dabei sowohl auf die Annahme, dass die mit CALP assoziierten eher schulischen bzw. akademischen Inhalte kognitiv komplexer sind als die mit BICS assoziierten alltagsnahen Inhalte als auch auf die Annahme, dass CALP in dekontextualisierten Situationen gebraucht wird. Kritiker der BICS/CALP-Unterscheidung sehen die kognitive Komplexität von Inhalten als abhängig vom thematischen Vorwissen: für Personen mit höherem Vorwissen sind Inhalte weniger komplex als für Personen mit geringem Vorwissen. Dies trifft auch auf eine Vielzahl nichtschulischer Themen zu, die laut J. Cummins‘ Definition nicht unter CALP fallen sollten

(Aukerman, 2007). Der zweite Kritikpunkt bezieht sich nicht direkt auf die Theorie von Cummins, sondern eher auf eine Auslegung dieser Theorie, die CALP als eine Voraussetzung für Lernerfolg sieht. Laut Aukerman (2007) sind Lehrpersonen bisweilen der Ansicht, dass Kompetenzerwerb nur eingeschränkt stattfinden kann, wenn die Lernenden nicht über ein Mindestmaß von CALP verfügen. Eine mangelnde Beherrschung von CALP wird hierbei von Lehrpersonen häufig daran festgestellt, dass die Kinder nicht in der Lage sind, sich an die sozialen Normen der schulischen Kommunikation zu halten. Daher sind nach Meinung der Kritiker schulische Interaktionen nicht dekontextualisiert. Vielmehr bietet die Lehr-Lern-Situation selbst einen Kontext, der allerdings von Lehrpersonen anders interpretiert werden kann als von Lernenden. Aukerman (2007) beschreibt, dass unterrichtliche Situationen häufig mit spezifischen, meist nicht explizierten Erwartungen an kindliche Sprache verbunden sind, wie etwa, dass die Kinder abstrakte Erklärungen und Anweisungen verstehen können. Kinder, die mit diesen Erwartungen nicht vertraut sind, können daher Probleme haben, diese zu erfüllen. Diese Probleme könnten dann dazu führen, dass den Kindern eine mangelnde Beherrschung von CALP zugeschrieben wird, die sich hinderlich auf den Lernerfolg auswirkt. Diese Interpretation von CALP als notwendige Voraussetzung für erfolgreichen Kompetenzerwerb kann zu einer defizitorientierten Beschreibung von Kindern führen, deren Kommunikationsstil nicht den sozialen Normen schulischer Kommunikation entspricht (MacSwan, 2000). Eine Benachteiligung von Zweitsprachlernenden käme somit durch mangelnde Passung der von den Schülerinnen und Schülern beherrschten Kommunikationsstile mit den in schulischen Interaktionskontexten verlangten sprachlichen Formen zustande: "[...] children command a range of registers even when they first come to school, regardless of what their home language is. However, children control different registers, depending on their home experiences, and the registers they control are not always the ones that are valued in school" (Schleppegrell, 2012a, S. 411).

Eine weitere Konzeption von Bildungssprache basiert auf einer Beschreibung bildungssprachlicher Funktionen im Unterricht. Die systemische funktionale Linguistik nach Halliday bietet einen theoretischen Rahmen für eine solche Beschreibung und wurde in mehreren Arbeiten von Mary Schleppegrell für Analysen des bildungssprachlichen Registers genutzt (Schleppegrell, 2001, 2004, 2007, 2012b). Ein Register bezeichnet „the constellation of lexical and grammatical features that realizes a particular situational context“ (Schleppegrell, 2001, S. 18). Der Kontext wird dabei durch die Merkmale „Thema des Textes“ (field), “Beziehung zwischen Sender und Empfänger“ (tenor) sowie „situationsspezifische Erwartungen an die Textorganisation“ (mode) beschrieben. Die systemische funktionale Linguistik betont die Funktion und die soziale Einbettung von Sprache in Interaktionskontexten. Dieser An-

satz geht davon aus, dass die Wahl der sprachlichen Mittel in kommunikativen Situationen bestimmte Funktionen erfüllen muss. Schleppegrell (2001, 2004) stellt dabei heraus, dass schulische Interaktionen weder von sich aus komplexer oder expliziter noch dekontextualisierter sind als alltägliche Interaktionen. Dennoch werden in der Schule andere sprachliche Mittel gebraucht als in alltäglichen Interaktionen, da sich die zu erfüllenden sprachlichen Funktionen zwischen schulischen Situationen und Freizeitsituationen unterscheiden. Eine dem schulischen Umfeld zuzuordnende Funktion ist beispielsweise „sich als eine wissende Person präsentieren“, wohingegen die Funktion „Mitteilen von persönlichen Erlebnissen“ eher den alltäglichen Interaktionen zuzuordnen ist (Schleppegrell, 2004). Auch Nagy und Townsend (2012) weisen in ihrer Untersuchung von bildungssprachlichem Wortschatz auf dessen Funktionalität hin: "Learning academic language is not learning new words to do the same thing that one could have done with other words; it is learning to do new things with language and acquiring new tools for these new purposes." (Nagy & Townsend, 2012, S. 93). Ebenso wie in anderen Fächern ist auch in Mathematik Bildungssprache funktional und erweitert die Kommunikationsmöglichkeiten fachlicher Inhalte über die Alltagssprache hinaus: "Because concepts that mathematics construct are often difficult to articulate in ordinary language, mathematics symbolism has developed to express meanings that go beyond what ordinary language can express." (Schleppegrell, 2007, S. 141).

2.3.2 Familiäre und institutionelle Lerngelegenheiten für den Erwerb der Bildungssprache

In der Literatur werden verschiedene familiäre Hintergrundmerkmale und institutionelle Gegebenheiten diskutiert, die besonders mit der Entwicklung bildungssprachlicher Fähigkeiten in Beziehung stehen. Im Bezug auf das familiäre Umfeld werden vor allem die Merkmale sprachliches Anregungsniveau in der Familie, SES und Familiensprache als Einflussgrößen für bildungssprachliche Fähigkeiten angenommen (Genesee & Lindholm-Leary, 2012), wobei unter dem sprachlichen Anregungsniveau die Häufigkeit und Qualität von Lese- und Sprachaktivitäten im Elternhaus verstanden wird.

Es wird angenommen, dass insbesondere die Kommunikation im familiären Umfeld einen wesentlichen Einfluss auf die Entwicklung bildungssprachlicher Fähigkeiten hat (Huttenlocher, Waterfall, Vasilyeva, Vevea & Hedges, 2010; Leseman, Scheele, Mayo & Messer, 2007; McElvany, Becker & Lüdtke, 2009; Rindermann & Baumeister, 2015; Scheele, Leseman & Mayo, 2010). Kommunikative Situationen in der Familie bieten somit wichtige Lerngelegenheiten für den Erwerb und die Anwendung komplexerer sprachlicher Strukturen (vgl. Domenech & Krah, 2014; Hoff, 2003; Huttenlocher et al., 2010). Hier kann zwischen zwei Typen familiärer Kommunikation unterschieden werden. Zum einen können

alltägliche Konversationen, beispielsweise während der Mahlzeiten, den Kindern Gelegenheit bieten, bestimmte sprachliche Strukturen zu hören und zu verwenden. Zum anderen können gemeinsame Lese- und Sprachaktivitäten von Eltern und Kindern zusätzlich zur Entwicklung sprachlicher Fähigkeiten beitragen.

Die Ergebnisse längsschnittlicher Studien weisen darauf hin, dass sich alltägliche Konversationen zwischen Familien mit unterschiedlichem SES unterscheiden. Beispielsweise verwendeten Mütter mit einem hohen SES gegenüber ihren Kindern häufiger lexikalisch und syntaktisch komplexe sprachliche Strukturen als Mütter mit niedrigerem SES (z. B. Hoff, 2003; Huttenlocher et al., 2010). Diese Studien konnten auch zeigen, dass Vorschulkinder, deren Mütter häufiger komplexe sprachliche Strukturen verwendeten, die entsprechenden Strukturen früher erwarben. Außerdem wiesen Familien mit hohem SES ein höheres sprachliches Anregungsniveau auf, welches sich daran zeigte, dass in den Haushalten mehr Bücher und Bilderbücher vorhanden waren und häufiger gemeinsam gelesen wurde (z. B. Niklas & Schneider, 2013; Scheele et al., 2010; van Steensel, 2006). Auch für das sprachliche Anregungsniveau konnten positive Beziehungen zu komplexeren sprachlichen Fähigkeiten gezeigt werden (Leseman et al., 2007; Scheele, Leseman, Mayo & Elbers, 2012). Diese mit dem SES der Familien in Beziehung stehenden Unterschiede in beiden Typen familiärer Kommunikation geben Anlass zu der Annahme, dass sich die bildungssprachlichen Fähigkeiten von Schülerinnen und Schülern mit hohem SES und Schülerinnen und Schülern mit niedrigem SES unterscheiden.

Schülerinnen und Schüler mit nicht-deutscher Familiensprache wachsen häufig in Familien mit niedrigem SES auf (Haag et al., 2012; Segeritz, Walter & Stanat, 2010). Die mit niedrigem SES in Beziehung stehenden Unterschiede in der familiären Kommunikation könnten daher für Schülerinnen und Schüler mit nicht-deutscher Familiensprache zu geringeren Lerngelegenheiten für den Erwerb bildungssprachlicher Fähigkeiten führen. Unabhängig vom SES der Familien haben Schülerinnen und Schüler mit nicht-deutscher Familiensprache weniger Gelegenheit, in ihren Familien sprachliche Strukturen der deutschen Sprache kennenzulernen und anzuwenden. Daher kann davon ausgegangen werden, dass Schülerinnen und Schüler mit nicht-deutscher Familiensprache beim Erwerb bildungssprachlicher Fähigkeiten höhere Hürden zu überwinden haben (vgl. Aarts, Demir & Vallen, 2011). Die Beziehungen zwischen Familiensprache, SES und bildungssprachlichen Fähigkeiten werden in Teilstudie 1 in Anhang A (Kapitel 6) ausführlicher beschrieben.

Auf institutioneller Seite zeigt sich, dass im normalen Fachunterricht meist nur wenig Lerngelegenheiten für den Erwerb bildungssprachlicher Fähigkeiten bestehen (Ernst-Slavit &

Mason, 2011; Gogolin, 2009). Es wird außerdem davon ausgegangen, dass verschiedene bildungssprachliche Merkmale unterschiedlich häufig explizit im Unterricht thematisiert werden. Insbesondere die Fachsprache, also das Verstehen und Verwenden von Fachbegriffen des jeweiligen Unterrichtsfachs, sollte aufgrund der curricularen Verankerung deutlich häufiger zum Gegenstand des Unterrichts werden als andere, fächerübergreifend bedeutsame sprachliche Merkmale. Diese Annahme bestätigte sich beispielsweise in einer explorativen Untersuchung von gesprochener Sprache im Naturwissenschaftsunterricht in der 5. Klasse, in der Bailey et al. (2007) feststellten, dass fachspezifischer bildungssprachlicher Wortschatz von Lehrkräften häufiger erklärt wird als allgemeiner bildungssprachlicher Wortschatz. Auch in Mathematik ist davon auszugehen, dass fachspezifischer Wortschatz im Unterricht explizit thematisiert und geübt wird (Pierce & Melena, 2009; Verboom, 2008), wohingegen anderen bildungssprachlichen Merkmalen weniger Aufmerksamkeit gewidmet wird. Insgesamt liegen jedoch nur wenige systematische Untersuchungen der bildungssprachlichen Merkmale von typischen Lehr-Lern-Situationen vor, sodass bislang nicht vollständig geklärt ist, ob andere bildungssprachliche Merkmale im Unterricht unterschiedlich häufig explizit angesprochen werden.

Schulische Lerngelegenheiten zum Erwerb von Bildungssprache sind insbesondere für Schülerinnen und Schüler relevant, die wenig familiäre Vorerfahrungen mit dem Register der Bildungssprache haben, also beispielsweise für Zweitsprachlernende oder für Kinder aus Familien mit niedrigem sozioökonomischem Status (Schleppegrell, 2001). Die schulische Vermittlung von Bildungssprache kann demnach auch als Frage der Chancengerechtigkeit betrachtet werden (z. B. Ehlich, 2013; Schleppegrell, 2012b). Es gibt Hinweise darauf, dass sich ein Angebot von angemessenen Lerngelegenheiten in der Schule positiv auf die bildungssprachlichen Fähigkeiten von Schülerinnen und Schülern auswirkt. Beispielsweise konnten Gámez und Lesaux (2012) positive Einflüsse der sprachlichen Komplexität von Lehreräußerungen auf den bildungssprachlichen Wortschatz von Schülerinnen und Schülern in der sechsten Klasse nachweisen. Interessanterweise bestand für alle Schülerinnen und Schüler eine positive Beziehung zwischen der Komplexität des Wortschatzes in den Lehreräußerungen und dem Zuwachs im bildungssprachlichen Wortschatz, für die Komplexität der Grammatik der Lehreräußerungen bestand diese Beziehung jedoch nur für einsprachig englische Schüler und für Zweitsprachlernende mit mindestens mittlerem Sprachniveau in der Unterrichtssprache. Zweitsprachlernende mit niedrigem Sprachniveau konnten hingegen nicht von grammatisch komplexen Äußerungen des Lehrers profitieren. Die Autoren interpretieren diesen Befund dahingehend, dass zu komplexer Input von Schülern mit niedrigem Sprachniveau ausgefiltert wird. Die Quantität des Inputs scheint dabei von

geringerer Bedeutung zu sein. Daraus lässt sich schließen, dass ein Mangel an außerschulischen Lerngelegenheiten durch einen den Sprachkenntnissen von Zweitsprachlernenden angemessen komplexen sprachlichen Input in Bildungsinstitutionen aufgefangen werden könnte. Diese schulischen Lerngelegenheiten scheinen allerdings im regulären Unterricht nur selten in ausreichendem Umfang gegeben zu sein, um die bildungssprachlichen Fähigkeiten von Zweitsprachlernenden nachhaltig zu fördern (Ernst-Slavit & Mason, 2011; Gogolin, 2009).

Die beschriebenen Befunde zu geringeren familiären Lerngelegenheiten für bildungssprachliche Fähigkeiten in Familien mit niedrigem SES oder mit nicht-deutscher Familiensprache einerseits und dem Mangel an angemessenen institutionellen Lerngelegenheiten, in denen Bildungssprache funktional ist, andererseits deuten darauf hin, dass bildungssprachliche Anforderungen von Testaufgaben Zweitsprachlernenden besondere Probleme bereiten könnten.

2.3.3 Bildungssprachliche Merkmale von Testaufgaben

Beruhend auf den theoretischen Grundlagen des Konstrukts Bildungssprache (vgl. Abschnitt 2.3.1) lassen sich Merkmale beschreiben, die den bildungssprachlichen Gehalt von Texten angeben. Die meisten Beschreibungen bildungssprachlicher Textmerkmale stammen aus dem englischsprachigen Raum (z. B. Abedi et al., 1997; Bailey et al., 2007; Nagy & Townsend, 2012; Prediger, 2013; Schleppegrell, 2001, 2004, 2007). Im deutschsprachigen Raum verbreitete Beschreibungen bildungssprachlicher Merkmale basieren auf diesen Vorbildern aus dem englischsprachigen Raum und wurden um spezifische Merkmale der deutschen Sprache, wie beispielsweise Präfixverben und Partikelverben, ergänzt (vgl. Berendes et al., 2013; Eckhardt, 2008; Heppt et al., 2012; Heppt et al., 2014). Insgesamt wurde eine Vielzahl von Indikatoren für die sprachliche Komplexität von Testaufgaben und Unterrichtstexten verwendet. Während einige wenige Studien die bildungssprachlichen Merkmale gesprochener Sprache von jüngeren Kindern und ihren Bezugspersonen beschreiben (Aarts et al., 2011; Demir-Vegter, Aarts & Kurvers, 2014; Leseman et al., 2007; Scheele et al., 2010), konzentrieren sich die meisten Arbeiten auf die Merkmale bildungssprachlicher Texte oder strukturierter Unterrichtskonversationen (z. B. Abedi et al., 1997; Bailey et al., 2007; Bailey & Huang, 2011; Nagy & Townsend, 2012; Prediger, 2013; Schleppegrell, 2001, 2004, 2007). In der bisherigen Forschung wurde ein Schwerpunkt auf die Operationalisierung des bildungssprachlichen Wortschatzes gelegt, wohingegen grammatische oder textorganisatorische Merkmale weniger häufig beschrieben und untersucht wurden (vgl. Uccelli, Galloway, Barr, Meneses & Dobbs, 2015). Die bisher beschriebenen Merkmale wurden in verschiedenen Reviews systematisiert (Scarcella, 2003; Snow & Uccelli, 2009). Diese Überblicksarbeiten

deuten jedoch an, dass bislang noch nicht geklärt werden konnte, in welcher Beziehung die einzelnen bildungssprachlichen Merkmale zueinander stehen (Snow & Uccelli, 2009).

Die Definition der bildungssprachlichen Merkmale von Texten lehnt sich im Allgemeinen eng an die Beschreibung des schulischen Registers (Schleppegrell, 2001, 2004, 2007, 2012b) und seiner charakteristischen linguistischen Merkmale an. Mary Schleppegrell zeigte in diesen Arbeiten auf, wie sich das schulische Register in Wortschatz, Grammatik und Textorganisation von anderen Registern, insbesondere vom Register alltäglicher Konversationen, unterscheidet. Wie in Abschnitt 2.3.1 beschrieben, erfüllen die sprachlichen Mittel des schulischen Registers für dieses Register spezifische Funktionen. Laut Schleppegrell (2001) gilt vor allem die lexikalische Dichte, also der Anteil von Substantiven am Text, als eines der bestimmenden Merkmale von Bildungssprache. Diese wird vor allem durch Nominalisierungen, lange Nominalphrasen und eingebettete Sätze erreicht. Mit Hilfe dieser sprachlichen Merkmale werden die einzelnen Sätze innerhalb eines schulischen Textes miteinander verbunden, um einen kohärenten Text zu formen. Demgegenüber weisen alltägliche Konversationen eine geringere lexikalische Dichte auf. Die Kohärenz der Konversation wird eher durch nonverbale bzw. prosodische Elemente hergestellt. Auf der grammatischen Ebene werden im schulischen Register beispielsweise mehr *unterschiedliche* Konjunktionen zur Verbindung von Sätzen benutzt als im alltagssprachlichen Kontext, was zu einer stärkeren hierarchischen Struktur schulischer Texte beiträgt.

Die im funktional-linguistischen Ansatz herausgearbeiteten Merkmale der Bildungssprache (Schleppegrell, 2001, 2004, 2007, 2012b) wurden von Bailey, Butler, LaFramenta und Ong (2004) aufgegriffen und anhand umfangreicher, systematischer Analysen der sprachlichen Komplexität von naturwissenschaftlichen, sozialwissenschaftlichen und mathematischen Unterrichtswerken um weitere Merkmale ergänzt, wobei für den Bereich Mathematik nur Textaufgaben betrachtet wurden. Aus diesen Überarbeitungen resultiert ein umfassendes Ratingschema für die sprachliche Komplexität von Unterrichtsmaterialien (Bailey et al., 2007). Die Autoren unterscheiden deskriptive, lexikalische, grammatische und diskursive Maße. Deskriptive Maße beziehen sich auf die Textlänge—gemessen durch die Anzahl an Wörtern und Sätzen—sowie die mittlere Satzlänge. Auf lexikalischer Ebene werden lange und wenig gebräuchliche Wörter, die lexikalische Diversität des Textes sowie die Anzahl allgemeiner und fachspezifischer bildungssprachlicher Wörter erfasst. Grammatische Merkmale umfassen die Anzahl verschiedener Satzarten, Anzahl und durchschnittliche Länge von Nominal- und Präpositionalphrasen, Passivkonstruktionen, Partizipien und Nominalisierungen.

Auch das für die vorliegende Arbeit verwendete Ratingschema orientiert sich an den von Bailey et al. (2007) vorgestellten Merkmalen. Dieses Schema wurde gewählt, da es sich dabei um eine der umfassendsten konkreten Beschreibungen bildungssprachlicher Merkmale handelt, welches auch für Testaufgaben geeignet ist. Es hat sich außerdem bereits in entsprechenden empirischen Arbeiten bewährt (z. B. Wolf & Leon, 2009) und bildete die Basis für im deutschsprachigen Raum verwendete Kodierschemata (vgl. Berendes et al., 2013; Eckhardt, 2008; Heppt et al., 2012; Heppt et al., 2014). Eine genauere Beschreibung der für die vorliegende Arbeit kodierten sprachlichen Merkmale inklusive entsprechender Beispiele findet sich in den Teilstudien 1 und 2 in Anhang A (Kapitel 6) bzw. in Anhang B (Kapitel 7).

Auf lexikalischer Ebene lassen sich zusätzlich zu den von Bailey et al. (2007) beschriebenen Merkmalen die Anzahl der Wörter mit lateinischen oder griechischen Wurzeln, die Anzahl abgeleiteter Wörter (v.a. durch Prä- und Suffixe) sowie ein im Vergleich zur Alltagssprache höherer Anteil von Substantiven, Adjektiven und Präpositionen als Indikatoren für sprachliche Komplexität heranziehen (Nagy & Townsend, 2012).

Auf theoretischer Ebene ist bislang nicht abschließend geklärt, inwiefern sich das Konstrukt der Bildungssprache im Verlauf der Schulzeit verändert und welche Merkmale in verschiedenen Klassenstufen als zentral für die Bildungssprache gelten sollten (Snow & Uccelli, 2009). Empirische Arbeiten zu linguistischen Merkmalen von Texten deuten darauf hin, dass in der Schule verwendete Texte im Verlauf der Schulzeit an sprachlicher Komplexität zunehmen (Deane, Sheehan, Sabatini, Futagi & Kostin, 2006; Graesser, McNamara & Kulikowich, 2011). Diese zunehmende Komplexität zeigt sich sowohl im Bereich des Wortschatzes als auch im Bereich der Grammatik. Entsprechend umfassende vergleichende Untersuchungen für Aufgaben aus den Sachfächern stehen jedoch noch aus. Erste Studien in diese Richtung wurden von Wolf und Leon (2009) und von Shaftel, Belton-Kocher, Glasnapp und Poggio (2006) durchgeführt. Shaftel et al. (2006) untersuchten die sprachliche Komplexität der in Kansas verwendeten staatlichen Mathematiktests der Klassenstufen 4, 7 und 10 und konnten einen Anstieg der sprachlichen Komplexität von Mathematikaufgaben über die untersuchten Klassenstufen feststellen. Wolf und Leon (2009) verglichen die sprachlichen Anforderungen der standardisierten Mathematik- und Naturwissenschaftstests aus drei US-Bundesstaaten für die Klassenstufen 4, 5, 7 und 8. Die Befunde deuten darauf hin, dass es zwar insgesamt einen Anstieg der sprachlichen Komplexität der Testaufgaben gibt, die Varianz in der sprachlichen Komplexität der Testverfahren aber auch innerhalb einer Klassenstufe beträchtlich ist.

2.4 Befunde zum Einfluss sprachlicher Aufgabenmerkmale auf Testleistungen

Sprachliche Anforderungen von Testaufgaben in Sachfächern wie Mathematik oder Naturwissenschaften können als konstruktirrelevante Eigenschaften von Aufgaben gesehen werden, durch die Zweitsprachlernende potenziell benachteiligt sind (vgl. Abedi, 2002; Young, 2009). Auch die *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999) weisen darauf hin, dass die Sprachfähigkeit in Testverfahren für Sachfächer implizit mitgemessen wird. Daher können Sprachfähigkeiten in Mathematiktests als zusätzliche Dimension gesehen werden, deren Nicht-Modellierung sich auf Zweitsprachlernende stärker auswirken könnte als auf Erstsprachlernende (vgl. Ackerman, 1992), was zu differenziellem Itemfunktionieren und damit einer eingeschränkten Fairness der Tests führen könnte (vgl. Kapitel 2.2.1). Die Auswirkungen dieser versteckten Mehrdimensionalität von Mathematiktests auf verschiedene Teststatistiken wie Reliabilität, Anteil von fehlenden Werten und differenzielles Itemfunktionieren wurden in einer Vielzahl von empirischen Studien geprüft, deren Ergebnisse jedoch uneinheitlich sind.

Verschiedene Studien setzten think-aloud Protokolle ein, um zu überprüfen, ob sich die Antwortprozesse von Erstsprachlernenden und Zweitsprachlernenden unterscheiden. In diesen Studien wird untersucht inwieweit Zweitsprachlernende Testaufgaben aufgrund sprachlich bedingter Verständnisschwierigkeiten überproportional häufig falsch beantworten (Ercikan et al., 2010; Martiniello, 2008; Noble et al., 2012; Young et al., 2014). Die Befunde dieser Studien sind uneindeutig. Noble et al. (2012) fanden, dass die sprachlichen Merkmale von Aufgaben eines Naturwissenschaftstests bei ELLs zu Problemen führten, die Aufgaben richtig zu beantworten, obwohl sie über das entsprechende Wissen verfügen. Martiniello (2008) konnte ähnliche spezifische Probleme von spanischsprachigen ELLs beim Verständnis von Mathematikaufgaben beobachten. Aus einer deutschsprachigen Studie von Prediger, Renk, Büchter, Gürsoy und Benholz (2013) lässt sich ableiten, dass Zweitsprachlernenden vor allem das Erkennen mathematischer Zusammenhänge Probleme bereitet. Allerdings zeigte die Studie von Noble et al. (2012) ebenfalls eine beträchtliche Anzahl von Verständnisproblemen für Schülerinnen und Schüler mit englischer Erstsprache, die auch in dieser Gruppe mit Problemen bei der Aufgabenbearbeitung einhergingen. Ebenso kamen Young et al. (2014) in einer neueren Studie zu dem Schluss, dass vom Schüler berichtetes Verstehen von Mathematikaufgaben nicht zwingend mit der korrekten Lösung dieser Aufgaben in Zusammenhang steht. Die Autoren fanden zwar heraus, dass sprachlich vereinfachte Varianten von Testaufgaben in den Bereichen Mathematik und Naturwissenschaften zwar tendenziell besser verstanden wurden, die Lösungswahrscheinlichkeit der Aufgaben unterschied sich jedoch nicht signifikant zwischen sprachlich komplexen und

sprachlich vereinfachten Aufgaben. Diese Befunde zeigen, dass Zweitsprachlernende spezifische Schwierigkeiten beim Verständnis mathematischer und naturwissenschaftlicher Aufgaben zu haben scheinen. Es ist jedoch nicht klar, inwieweit sich diese sprachlich bedingten Schwierigkeiten auf die Lösung der Aufgaben auswirken und ob eine sprachliche Vereinfachung der Aufgaben zu höheren Lösungswahrscheinlichkeiten führt.

Unterschiede im Einfluss sprachlicher Fähigkeiten zwischen Erstsprachlernenden und Zweitsprachlernenden auf die Bearbeitung von Mathematikaufgaben könnten ferner dazu führen, dass sich die Struktur des Tests für Erstsprachlernende und für Zweitsprachlernende unterscheidet. Ein solcher Effekt ließe sich daran feststellen, dass die Reliabilität des Tests für Zweitsprachlernende niedriger ausfällt als für Erstsprachlernende, was sich als eine geringere Messgenauigkeit des Tests für Zweitsprachlernende interpretieren lässt. Dies würde darauf hindeuten, dass der Mathematiktest in der Gruppe der Zweitsprachlernenden weniger genau zwischen Personen mit unterschiedlicher Mathematikfähigkeit differenzieren kann. Neben einem stärkeren direkten Einfluss sprachlicher Fähigkeiten bei der Bearbeitung der Mathematikaufgaben könnte eine geringere Messgenauigkeit des Tests in der Gruppe der Zweitsprachlernenden auch auf einen höheren Anteil fehlender Testantworten in dieser Gruppe zurückzuführen sein. Insbesondere fehlende Antworten am Ende des Tests könnten darauf hindeuten, dass es den Testteilnehmern nicht möglich war, den Test in der vorgegebenen Zeit vollständig zu bearbeiten und das Testergebnis somit neben der Mathematikleistung auch die Geschwindigkeit der Testbearbeitung widerspiegelt. Für amerikanische Schulleistungstests konnten geringfügig niedrigere Reliabilitäten und höhere Anteile fehlender Werte für English Language Learners (ELLs) als für monolingual englischsprachige Schülerinnen und Schüler gezeigt werden (Abedi, 2002; Abedi et al., 1997; Young et al., 2008). Die geringeren Reliabilitäten deuten darauf hin, dass die Testverfahren für ELLs mit tendenziell größeren Messfehlern verbunden sind. Aus dem höheren Anteil fehlender Werte lässt sich schließen, dass die Verarbeitung der Testaufgaben bei ELLs möglicherweise mehr kognitive Ressourcen in Anspruch nimmt und dadurch Zeitprobleme bei der Aufgabenbearbeitung auftreten können. Eine größere Beanspruchung kognitiver Ressourcen könnte aus den in den think-aloud Studien gefunden Verständnisschwierigkeiten resultieren und somit auf Unterschiede in den Antwortprozessen zwischen Erstsprachlernenden und Zweitsprachlernenden zurückzuführen sein.

In der bisherigen Forschung zu den Effekten sprachlicher Anforderungen von Testaufgaben auf die Messung von Schülerfähigkeiten kommt DIF-Analysen ein großer Stellenwert zu, da sich mit dieser Art der Analysen *item bias* gut von *item impact* isolieren lässt (vgl. Ackerman, 1992). Verglichen mit Analysen zu Unterschieden in der Testreliabilität und den der Testbe-

arbeitung zugrundeliegenden Antwortprozessen wurden DIF-Analysen deutlich häufiger durchgeführt, sodass die meisten Erkenntnisse über die Fairness von Testverfahren auf DIF-Analysen der Items beruhen. Verschiedene Studien untersuchten die Zusammenhänge zwischen globalen Ratings sprachlicher Komplexität oder einzelnen bildungssprachlichen Merkmalen und DIF zu Ungunsten von Zweitsprachlernenden (Abedi & Lord, 2001; Abedi et al., 1997; Martiniello, 2009; Prediger et al., 2013; Shaftel et al., 2006; Wolf & Leon, 2009). Die Ergebnisse sind jedoch uneindeutig. Beispielsweise sagte Martiniello (2009) die für einen englischsprachigen Mathematiktest für die vierte Klasse festgestellten DIF-Werte zu Ungunsten von Zweitsprachlernenden durch eine globale Einschätzung der konstruktirrelevanten sprachlichen Komplexität der Mathematikaufgaben vorher. In dieser Untersuchung wiesen sprachlich komplexe Aufgaben signifikant mehr DIF auf, was sich dahingehend interpretieren lässt, dass die sprachlich komplexen Mathematikaufgaben für Zweitsprachlernende differenziell valide waren. Andere Studien konnten diese Ergebnisse jedoch nicht bestätigen und fanden keine Zusammenhänge zwischen der sprachlichen Komplexität von Testaufgaben und DIF zuungunsten von Zweitsprachlernenden (Hickendorff, 2013; Shaftel et al., 2006). Neben den Zusammenhängen globaler Einschätzungen sprachlicher Komplexität wurde auch untersucht, inwiefern einzelne bildungssprachliche Merkmale unterschiedlich starke Beziehungen zu differenziellem Itemfunktionieren aufweisen. Dabei wurde häufig angenommen, dass sich vor allem diejenigen Merkmale benachteiligend auswirken sollten, die im Unterricht eher selten explizit thematisiert werden. Dementsprechend sollte beispielsweise mathematikspezifische sprachliche Anforderungen weniger benachteiligend wirken, da fachspezifischer Wortschatz im Unterricht explizit thematisiert und geübt wird (Pierce & Melena, 2009; Verboom, 2008). Der Forschungsstand zur Benachteiligung von Zweitsprachlernenden durch *verschiedene* bildungssprachliche Merkmale ist in Teilstudie 2 in Anhang B (Kapitel 7) ausführlich beschrieben und wird daher hier nicht dargestellt.

Die Effekte sprachlicher Komplexität von Testaufgaben auf die Lösungswahrscheinlichkeit dieser Aufgaben für Zweitsprachlernende wurden auch in einer Reihe von experimentellen Studien geprüft, in der Testaufgaben anhand bestimmter Kriterien sprachlich vereinfacht wurden (für eine Zusammenfassung gängiger Kriterien siehe Abedi, 2009; Kopriva, 2008). Sprachlich vereinfachte Versionen von Testaufgaben in Sachfächern werden in den USA häufig als angepasste Testversionen für Zweitsprachlernende angeboten. Diese sogenannten Testakkommodationen sollen dazu dienen, die Leistungsmessung in Sachfächern für Zweitsprachlernende besser mit der Leistungsmessung für Erstsprachlernende vergleichbar zu machen, indem konstruktirrelevante sprachliche Barrieren soweit wie möglich reduziert werden. Hierbei ist zu beachten, dass sich die Vereinfachung nur auf *konstruktirrelevante*

sprachliche Aspekte beziehen darf, um den akkommodierten Schülern keinen Vorteil bei der Testbearbeitung zu geben. Ansonsten wäre die Vergleichbarkeit der Testergebnisse gefährdet. Gute Akkommodationen sollten also zu einem Interaktionseffekt führen: Schülerinnen und Schüler, die Akkommodationen benötigen, sollten davon profitieren; bekommen hingegen Schülerinnen und Schüler Akkommodationen, die diese nicht benötigen, sollten sich ihre Testleistungen nicht (oder nicht im gleichen Ausmaß) verbessern (Lynn S. Fuchs, Fuchs, Eaton, Hamlett & Karns, 2000; Sireci, Scarpati & Li, 2005).

Die bisherigen Befunde zu sprachlicher Vereinfachung von Testaufgaben wurden in drei Metaanalysen zusammengefasst (Kieffer et al., 2012; Li & Suen, 2012; Pennock-Roman & Rivera, 2011), die in Teilstudie 3 in Anhang C (Kapitel 8) umfassend beschrieben sind. Insgesamt lässt sich sagen, dass die Fairness der Leistungsmessung für Zweitsprachlernende durch sprachliche Vereinfachung leicht verbessert werden konnte. Allerdings stehen bei dieser Aussage wenige erfolgreiche sprachliche Vereinfachungen (Aguirre-Muñoz, 2000; Sato, Rabinowitz, Gallagher & Huang, 2010) einer größeren Anzahl von Studien gegenüber, die nur geringe oder keine Effekte der sprachlichen Vereinfachung zeigen konnten (z. B. Abedi & Gándara, 2006; Hofstetter, 2003; Johnson & Monroe, 2004).

Aus den Ergebnissen sowohl korrelativer als auch experimenteller Studien zum Einfluss konstruktirrelevanter sprachlicher Komplexität von Testaufgaben bestehen Hinweise darauf, dass sich der Einfluss dieser Komplexität innerhalb der Zweitsprachlernenden nach dem Sprachniveau der Schülerinnen und Schüler unterscheiden kann. Beispielsweise fanden Wolf und Leon (2009) in einer Untersuchung von Mathematik- und Naturwissenschaftstests verschiedener US-Bundesstaaten höhere DIF-Werte und ausgeprägtere Korrelationen der DIF-Werte mit den sprachlichen Anforderungen der Aufgaben für ELLs mit niedrigen Sprachkenntnissen im Gegensatz zur Gesamtgruppe der ELLs. Die Korrelation der DIF-Werte mit den mathematikspezifischen sprachlichen Anforderungen der Aufgaben fiel jedoch für ELLs mit niedrigen Sprachkenntnissen etwa gleich hoch aus wie für die Gesamtgruppe der ELLs. Einige Studien konnten außerdem zeigen, dass DIF-Effekte für Kinder mit schwachen sprachlichen Leistungen stärker mit sprachlichen Eigenschaften der Items in Beziehung stehen als DIF-Effekte für Zweitsprachlernende (Hickendorff, 2013; Prediger et al., 2013). Metaanalytische Befunde zu experimentellen Studien deuten darauf hin, dass vor allem diejenigen Schülerinnen und Schüler von sprachlichen Vereinfachungen von Testaufgaben profitieren können, die über ein mittleres Kompetenzniveau in der Testsprache verfügen (Pennock-Roman & Rivera, 2011). Die Effekte der Vereinfachung fielen für Schülerinnen und Schüler mit geringen Kompetenzen in der Testsprache oder für Schülerinnen und Schüler mit hohen Kompetenzen in der Testsprache dementsprechend geringer aus als für

Schülerinnen und Schüler mit mittleren Kompetenzen in der Testsprache. Aus diesen Befunden lässt sich ableiten, dass insbesondere die Kenntnisse in der Testsprache einen moderierenden Einfluss auf die Benachteiligung von Zweitsprachlernenden durch sprachlich komplexe Aufgaben haben sollten. Allerdings beruhen diese Befunde bislang auf wenigen Einzelstudien, die in der Sekundarstufe durchgeführt wurden (Aguirre-Muñoz, 2000; Sato et al., 2010).

3 Forschungsfragen

In der vorliegenden Arbeit soll untersucht werden, inwiefern die sprachliche Komplexität von Mathematikaufgaben einen benachteiligenden Einfluss auf die Erfassung der Mathematikleistung von Schülerinnen und Schülern mit nicht-deutscher Familiensprache ausübt. Zu diesem Zweck wird zunächst der Frage nachgegangen, inwieweit Mathematikaufgaben in der Grundschule Schülerinnen und Schüler mit nicht-deutscher Familiensprache benachteiligen und somit zu differenziell validen Aussagen über die Mathematikleistung dieser Gruppe führen können. Daran anschließend wird untersucht, ob sich die spezifischen Leistungsnachteile von Schülerinnen und Schülern mit nicht-deutscher Familiensprache durch die sprachlichen Merkmale der Aufgaben erklären lassen. Damit greift die vorliegende Arbeit eine Annahme auf, die schon seit längerer Zeit in der öffentlichen und wissenschaftlichen Diskussion existiert (Abedi, 2002; Abedi et al., 1997; Berendes et al., 2013; Butler, Stevens & Castellon, 2007; Eckhardt, 2008; Gogolin, 2009; Gogolin & Lange, 2011; Grießhaber, 2011; Rösch & Paetsch, 2011; Schründer-Lenzen, 2008). Obwohl der Einfluss sprachlicher Merkmale auf das Lösen von Mathematikaufgaben vor allem im amerikanischen Sprachraum bereits mehrfach untersucht wurde, blieb bislang unklar, wie stark die sprachlichen Anforderungen der Aufgaben die Fairness der Tests für Zweitsprachlernende einschränken. Einige dieser Studien weisen auf eine gering ausgeprägte Beeinträchtigung der Reliabilität sowie auf differenzielles Itemfunktionieren durch sprachlich komplexe Mathematikaufgaben hin (Abedi, 2002; Abedi & Lord, 2001; Abedi et al., 1997; Martiniello, 2009; Young et al., 2008). Das Ausmaß der Benachteiligung variiert über die betrachteten Testverfahren, ist aber insgesamt eher gering ausgeprägt. Entsprechende systematische Analysen von Mathematikaufgaben liegen für den deutschsprachigen Raum bislang nicht vor. Allerdings bestehen Hinweise darauf, dass sprachliche Merkmale von Mathematikaufgaben auch im Deutschen die Fairness von Schulleistungstests für Zweitsprachlernende einschränken könnten.

Um die sprachlichen Anforderungen der Aufgaben systematisch zu beschreiben, wird das Konstrukt der Bildungssprache verwendet. Bildungssprache bietet einen theoretischen Rahmen dafür, die mögliche Benachteiligung von Zweitsprachlernenden durch sprachlich komplexe Mathematikaufgaben zu erklären. Allerdings ist bislang unklar, inwiefern die angenommenen Benachteiligungen spezifisch für Zweitsprachlernende sind und wie sich diese Benachteiligung im Laufe der Schulzeit verändert. Bislang ist außerdem wenig über die Beziehungen der einzelnen bildungssprachlichen Merkmale untereinander und über ihre spezifischen Effekte auf Leistungsunterschiede zwischen Zweitsprachlernenden und Erstsprachlernenden in Mathematik bekannt.

Die vorliegende Arbeit soll dazu beitragen, Lücken sowohl in der Forschung zur differenziellen Validität von Mathematiktests in der Grundschule als auch in der Forschung zu Bildungssprache zu verringern. Aus den beschriebenen Befunden lassen sich drei Fragestellungen ableiten, die in der vorliegenden Arbeit verfolgt werden:

1. Bestehen Hinweise auf differenzielle Validität von Testaufgaben für Zweitsprachlernende?
2. Besteht ein Zusammenhang zwischen der bildungssprachlichen Komplexität von Aufgabentexten und der differenziellen Validität von Testaufgaben für Zweitsprachlernende?
3. Bestehen Unterschiede in den Zusammenhängen *verschiedener* bildungssprachlicher Merkmale der Aufgabentexte mit der differenziellen Validität von Testaufgaben für Zweitsprachlernende?

In der ersten Fragestellung soll untersucht werden, inwiefern Zweitsprachlernende in der Grundschule durch die in nationalen Schulleistungsstudien verwendeten Mathematikaufgaben differenziell benachteiligt werden. Das Ausmaß der Benachteiligungen gibt einen Hinweis darauf, ob die betrachteten nationalen Testverfahren geeignet sind, mathematische Kompetenzen in einer sprachlich heterogenen Schülerschaft fair zu erfassen und ob das wiederholt schlechte Abschneiden von Schülerinnen und Schülern nicht-deutscher Familiensprache in Mathematik teilweise auf Merkmale der Mathematikaufgaben zurückgeführt werden kann.

In der zweiten Fragestellung soll, aufbauend auf die erste Fragestellung, geprüft werden, ob sich eine differenzielle Benachteiligung von Zweitsprachlernenden durch bildungssprachliche Merkmale der Testaufgaben erklären lässt. Als notwendige Voraussetzung für die Hypothese, dass Zweitsprachlernende durch bildungssprachlich komplexe Aufgaben benachteiligt werden, muss ein Zusammenhang zwischen der bildungssprachlichen Komplexität der Aufgabentexte auf der einen Seite und dem Verständnis von Texten auf der anderen Seite gegeben sein. Dieser Zusammenhang muss dabei für Zweitsprachlernende stärker ausgeprägt sein als für Erstsprachlernende. Daher soll zunächst untersucht werden, inwiefern sich bildungssprachliche Merkmale von Texten für Zweitsprachlernende stärker auf das Textverstehen auswirken als für Erstsprachlernende. Anschließend soll geprüft werden, ob die bildungssprachlichen Merkmale von Mathematikaufgaben zu einer differenziellen Benachteiligung von Zweitsprachlernenden führen.

In der dritten Fragestellung wird ferner untersucht, inwiefern verschiedene bildungssprachliche Merkmale unterschiedliche Zusammenhänge mit differenziellen Leistungsnachteilen

von Zweitsprachlernenden aufweisen. Mit Hilfe dieser Befunde kann geprüft werden, ob sich die verschiedenen bildungssprachlichen Merkmale in ähnlicher Weise auf differenzielles Itemfunktionieren auswirken oder ob einzelne Merkmale in besonderem Maße benachteiligend wirken.

In den vier Teilstudien der vorliegenden Arbeit werden die drei Fragestellungen folgendermaßen aufgegriffen: In der ersten Teilstudie werden alle drei Fragestellungen für den Bereich Lesen untersucht. In dieser Teilstudie wird also geprüft, inwiefern bildungssprachlich komplexe Aufgaben für die betrachteten Schülergruppen überhaupt Probleme im Textverstehen verursachen. Im Rahmen dieser Studie wird auch verglichen, ob bildungssprachliche Merkmale von Texten Kinder aus Familien mit niedrigem sozioökonomischem Status (SES) vor ähnlich hohe Hürden stellen wie Zweitsprachlernende. In der zweiten Teilstudie werden die erste und die dritte Fragestellung für den Bereich Mathematik und für die Gruppen „Kinder mit deutscher Familiensprache“ und „Kinder mit nicht-deutscher Familiensprache“ untersucht. Diese Teilstudie fokussiert dabei die dritte Fragestellung und prüft, ob sich die Effekte der einzelnen bildungssprachlichen Merkmale voneinander unterscheiden. Die dritte und vierte Teilstudie beziehen sich auf die erste und zweite Fragestellung und untersuchen den Einfluss bildungssprachlicher Merkmale von Mathematikaufgaben, wobei nicht zwischen den einzelnen Merkmalen unterschieden wird. In der dritten Teilstudie werden diese Fragestellungen in einem experimentellen Ansatz geprüft, wohingegen der Fokus der vierten Teilstudie auf Klassenstufenunterschieden liegt.

In der ersten Teilstudie (*The Role of Academic-Language Features for Reading Comprehension of Language-Minority Students and Students From Low-SES Families*) wird zunächst betrachtet, inwiefern die bildungssprachlichen Anforderungen von Lesetestaufgaben mit spezifischen Leistungsnachteilen von Kindern mit niedrigem SES bzw. von Kindern mit nichtdeutscher Familiensprache in Beziehung stehen. Diese Teilstudie basiert auf der Hypothese, dass Schülerinnen und Schüler, die potenziell über geringere Lerngelegenheiten für Bildungssprache im Elternhaus verfügen, spezifische Probleme beim Verständnis bildungssprachlich komplexer Aufgaben haben. Dementsprechend wird untersucht, ob Zweitsprachlernende—die häufig in Familien mit niedrigem SES aufwachsen, aber zusätzlich über weniger familiäre Lerngelegenheiten für Bildungssprache in Deutsch verfügen—stärkere Probleme haben als Kinder aus Familien mit niedrigem SES. In der ersten Teilstudie werden neun Testaufgaben des IQB-Ländervergleichs Primarstufe 2011 (Stanat et al., 2012) aus dem Bereich Lesen verwendet, deren bildungssprachliche Komplexität nach einem auf den Arbeiten des CRESST basierenden Kodierschemas eingeschätzt wurde (Bailey et al., 2007). Dieses Kodierschema wurde für alle Teilstudien verwendet. Untersucht wird eine Teilstich-

probe von 19108 Schülerinnen und Schülern der vierten Jahrgangsstufe, die mindestens eine der neun Testaufgaben bearbeitet haben. Zur Erfassung des familiären Hintergrunds wurden die Familiensprache der Schülerinnen und Schüler sowie der berufliche Status ihrer Eltern genutzt.

In einer zweiten Teilstudie (*Second Language Learners' Performance in Mathematics: Disentangling the Effects of Academic Language Features*) wird das Zusammenspiel bildungssprachlicher Merkmale untersucht. Hierbei wird geprüft, inwiefern sich einzelne bildungssprachliche Merkmale spezifisch und gemeinsam auf die Leistungsnachteile von Kindern mit nicht-deutscher Familiensprache in Mathematik auswirken. Zu diesem Zweck wird eine Kommunalitätenanalyse der Daten der VERA-3 Erhebung 2010 in Berlin durchgeführt. Die 56 Mathematikitems des VERA-Tests wurden nach dem in Teilstudie 1 verwendeten Kodierschema eingeschätzt, wobei nur die Merkmale Textlänge, fachübergreifender bildungssprachlicher Wortschatz, mathematikspezifischer bildungssprachlicher Wortschatz, Anzahl der Nominalphrasen und Anzahl der Präpositionalphrasen in der Kommunalitätenanalyse betrachtet wurden. Bei VERA-3 handelt es sich um eine Vollerhebung aller Drittklässlerinnen und Drittklässler in Berlin, an der insgesamt 21618 Schülerinnen und Schüler teilnahmen.

Die dritte Teilstudie (*Linguistic Simplification of Mathematics Items: Effects for Language Minority Students in Germany*) untersucht experimentell, ob und gegebenenfalls unter welchen Bedingungen eine Reduktion der bildungssprachlichen Anforderungen von Mathematikaufgaben der vierten Klasse zu geringeren Leistungsnachteilen von Zweitsprachlernenden unter Kontrolle von Lesefähigkeit und SES führt. Die bisherige Forschung legt einen moderierenden Effekt des Sprachniveaus von Zweitsprachlernenden auf die Wirksamkeit von sprachlichen Vereinfachungen für diese Schülergruppe nahe. Allerdings beruhen diese Befunde bislang auf wenigen Einzelstudien, die in der Sekundarstufe durchgeführt wurden. In dieser Teilstudie wird daher untersucht, inwieweit Schülerinnen und Schüler mit mittlerer Sprachfähigkeit auch schon in der Grundschule besonders stark von sprachlichen Vereinfachungen profitieren. Zu diesem Zweck wurden 23 Mathematikitems aus dem Aufgabenpool für den IQB-Ländervergleich Primarstufe 2011 sprachlich vereinfacht und im IQB-Ländervergleich Primarstufe 2011 gemeinsam mit den originalen Items eingesetzt. Die Familiensprache der Schülerinnen und Schüler stellte die unabhängige Variable dar. Analysiert wurden die Daten einer Teilstichprobe von 17738 Schülerinnen und Schülern der vierten Jahrgangsstufe, welche ein Testheft bearbeiten, das unter anderem experimentell veränderte Aufgaben enthielt. Zur Untersuchung der Fragestellung wurde ein explanatorisches Item Response-Modell verwendet.

In der vierten Teilstudie (*Effects of Mathematics Items' Language Demands for Language Minority Students: Do They Differ Between Grades?*) wird untersucht, inwiefern sich die Effekte bildungssprachlicher Merkmale von Mathematikaufgaben auf die Mathematikleistungen von Kindern mit nicht-deutscher Familiensprache zwischen der dritten und der vierten Klassenstufe unterscheiden. Hierbei werden zwei verschiedene Datensätze unter Verwendung eines explanatorischen Item Response-Modells analysiert. Der erste Datensatz umfasst 26016 Schülerinnen und Schüler, die den Mathematiktest des IQB-Ländervergleichs Primarstufe 2011 bearbeitet haben. Der zweite Datensatz entstand im Rahmen der Normierungsstudie für den IQB-Ländervergleich Primarstufe 2011 und enthält Daten von 2919 Drittklässlerinnen und Drittklässlern. In beiden Studien wurden 126 gemeinsame Mathematikitems eingesetzt, deren sprachliche Anforderungen anhand des in den Teilstudien 1 und 2 verwendeten Kodierschemas eingeschätzt wurden. Die kodierten Merkmale wurden einer Hauptkomponentenanalyse unterzogen, um einen Faktor für die bildungssprachliche Komplexität der Aufgaben zu erhalten. Die Familiensprache der Schülerinnen und Schüler sowie die Anzahl der im Haushalt vorhandenen Bücher dienten als Indikatoren des familiären Hintergrunds.

4 Gesamtdiskussion

In der vorliegenden Arbeit wurde im deutschsprachigen Kontext geprüft, inwieweit sich die differenzielle Validität von Mathematikaufgaben für Kinder mit nicht-deutscher Familiensprache in der Grundschule durch die sprachlichen Merkmale dieser Aufgaben erklären lässt. Zu diesem Zweck wurde zunächst untersucht, wie hoch die itemspezifischen Leistungsnachteile für Kinder mit nicht-deutscher Familiensprache im Lesen und in Mathematik insgesamt sind. Hoch ausgeprägte Leistungsnachteile könnten auf differenzielle Validität der Testaufgaben für Kinder mit nicht-deutscher Familiensprache hinweisen. Ferner wurde der Frage nachgegangen, ob sich die sprachlichen Anforderungen der Testaufgaben insgesamt auf differenzielle Validität auswirken und ob verschiedene bildungssprachliche Merkmale unterschiedliche Zusammenhänge mit differenziellen Leistungsnachteilen aufweisen. Im Folgenden werden zunächst die zentralen Befunde der Arbeit zusammengefasst und in die bisherige Befundlage eingeordnet (Abschnitt 4.1). Anschließend werden die Grenzen der vorliegenden Arbeit aufgezeigt (Abschnitt 4.2) und die Implikationen der Arbeit für die Praxis der Testkonstruktion und für die Bildungspolitik werden diskutiert (Abschnitt 4.3). Abschließend zeigt Abschnitt 4.4 Perspektiven für weiterführende Forschung auf.

4.1 Diskussion der zentralen Befunde

4.1.1 Befunde zur differenziellen Validität der Testaufgaben für Zweitsprachlernende

In der ersten Fragestellung wurde untersucht, inwiefern Schülerinnen und Schüler mit nicht-deutscher Familiensprache durch die in nationalen Schulleistungsstudien verwendeten Mathematikaufgaben differenziell benachteiligt werden. Bezüglich dieser Fragestellung zeigte sich in allen vier Teilstudien übereinstimmend ein geringes Ausmaß differenzieller Leistungsnachteile von Schülerinnen und Schülern mit nicht-deutscher Familiensprache. Dies deutet darauf hin, dass die eingesetzten Testverfahren insgesamt als fair betrachtet werden können und geeignet sind, schulisch vermittelte Kompetenzen in einer sprachlich heterogenen Schülerschaft hinreichend gut zu erfassen.

Die Befunde der vorliegenden Arbeit decken sich mit Befunden aus dem amerikanischen Raum, in denen wiederholt differenzielle Benachteiligungen von Zweitsprachlernenden nachgewiesen werden konnten (Abedi & Lord, 2001; Martiniello, 2009; Wolf & Leon, 2009). Bemerkenswert ist jedoch, dass viele Studien nicht darauf eingehen, inwiefern diese differenziellen Benachteiligungen praktisch bedeutsam sind. In der Regel wird nur angegeben, ob die differenzielle Benachteiligung statistisch signifikant ist. Lediglich in der Studie von Martiniello (2009) wird die Varianz des DIF gegen Zweitsprachlernende angegeben (s^2 = 0.07). Dieser Wert spricht für eine sehr geringe Benachteiligung von Zweitsprachlernenden

und deckt sich mit dem Befund der vorliegenden Arbeit, dass die untersuchten Testverfahren als fair bezeichnet werden können. Auch im deutschsprachigen Raum wurden Nachteile von Zweitsprachlernenden bei sprachlastigen Aufgaben beschrieben (z. B. Duarte, Gogolin & Kaiser, 2011; Gogolin, 2009; Rösch & Paetsch, 2011), diese Beschreibungen sind jedoch zumeist anekdotisch. Die vorliegende Arbeit leistet somit einen wichtigen Beitrag zur Quantifizierung der differenziellen Nachteile von Zweitsprachlernenden in nationalen Schulleistungsstudien, die potenziell auf Merkmale der Testaufgaben zurückgeführt werden könnten.

4.1.2 Befunde zum Zusammenhang zwischen differenzieller Validität und den sprachlichen Anforderungen der Aufgaben

In der zweiten Fragestellung wurde geprüft, ob differenzielle Benachteiligungen von Schülerinnen und Schülern mit nicht-deutscher Familiensprache auf die sprachlichen Anforderungen der Testaufgaben zurückführbar sind. Trotz der insgesamt geringen Benachteiligung durch die Testaufgaben konnten in allen vier Teilstudien Zusammenhänge zwischen dem differenziellen Funktionieren von Items für Schülerinnen und Schüler mit nicht-deutscher Familiensprache und den bildungssprachlichen Anforderungen der Testaufgaben festgestellt werden. Diese Zusammenhänge unterschieden sich jedoch zwischen den Teilstudien. Insgesamt zeigte sich in den einzelnen Teilstudien, dass die bildungssprachlichen Anforderungen von Mathematiktestaufgaben einen geringen Teil des Leistungsnachteils von Kindern mit nicht-deutscher Familiensprache erklären konnten.

Die experimentell ausgerichtete Teilstudie 3 ergab keine substanziellen Effekte der sprachlichen Vereinfachung von Mathematiktestaufgaben auf die Testleistungen von Schülerinnen und Schülern der vierten Klassenstufe, weder von Schülerinnen und Schülern deutscher noch von Schülerinnen und Schülern mit nicht-deutscher Familiensprache. Zwar konnten leichte Verbesserungen von Schülerinnen und Schülern mit mittlerer Lesefähigkeit für die vereinfachten Aufgaben nachgewiesen werden, diese Effekte waren jedoch sehr gering und praktisch nicht bedeutsam. In Übereinstimmung mit metaanalytischen Befunden (Pennock-Roman & Rivera, 2011) zeigte sich in der vorliegenden Arbeit eine höhere Wirksamkeit sprachlicher Vereinfachung für Zweitsprachlernende mit mittleren Sprachkenntnissen. Allerdings fielen die Effekte der Vereinfachung insgesamt geringer aus, als aufgrund metaanalytischer Befunde aus dem englischsprachigen Raum erwartet werden konnte (Kieffer et al., 2012; Li & Suen, 2012; Pennock-Roman & Rivera, 2011).

Diese Abweichung zu bisherigen Studien könnte—neben dem offensichtlichen Unterschied der Sprache—auch darauf zurückzuführen sein, dass die in den Metaanalysen zusammengefassten Primärstudien hauptsächlich in der Sekundarstufe durchgeführt wurden. Die

sprachlichen Anforderungen von Lesetexten nehmen über die Schulzeit hinweg zu (Deane et al., 2006; Graesser et al., 2011). Daher erscheint es plausibel, dass auch die sprachlichen Anforderungen von Mathematikaufgaben für die Sekundarstufe höher sind als in der Primarstufe. Höhere sprachliche Anforderungen könnten außerdem zu stärkerer Benachteiligung von Schülerinnen und Schülern mit nicht-deutscher Familiensprache führen. Diese Benachteiligung würde sich empirisch als ein stärkerer Zusammenhang zwischen den sprachlichen Anforderungen der Aufgaben und dem differenziellen Itemfunktionieren zeigen. Die sprachliche Komplexität von Testaufgaben wurde bislang nur selten im Vergleich über mehrere Klassenstufen betrachtet. Wolf und Leon (2009) untersuchten die sprachlichen Anforderungen von elf Mathematik- und Naturwissenschaftstest für die Klassenstufen 4, 5, 7 und 8 aus drei U.S.-Staaten und konnten nur in einem Staat ein Ansteigen der bildungssprachlichen Anforderungen über die Klassenstufen feststellen. Zudem variierten die bildungssprachlichen Anforderungen innerhalb der betrachteten Klassenstufen beträchtlich über die Tests, sodass sich aus diesen Daten keine eindeutige Aussage über Unterschiede in der sprachlichen Komplexität der Aufgaben über die Klassenstufen ableiten lässt.

Die vorliegende Arbeit befasste sich über die einzelnen Teilstudien hinweg ebenfalls mit Unterschieden in den Zusammenhängen zwischen sprachlichen Anforderungen und differenziellem Itemfunktionieren zwischen Klassenstufen. In der vierten Klasse schien kein substanzieller Zusammenhang zwischen bildungssprachlichen Merkmalen der Aufgaben und differenzieller Benachteiligung von Schülerinnen und Schülern mit nicht-deutscher Familiensprache mehr zu bestehen (Teilstudie 4), wohingegen im Lesen noch differenzielle Benachteiligung festgestellt werden konnte (Teilstudie 1). In Teilstudie 4 zeigten sich außerdem in Mathematik stärkere Zusammenhänge für die dritte als für die vierte Klassenstufe. Dieser Befund wurde auch von Teilstudie 2 gestützt, bei der sich in der dritten Klassenstufe substanzielle Zusammenhänge zwischen dem differenziellen Itemfunktionieren von Mathematikitems und den sprachlichen Anforderungen dieser Items nachweisen ließen.

Bisherige Studien konnten noch keine reliablen Unterschiede in den Zusammenhängen zwischen sprachlichen Anforderungen und differenziellem Itemfunktionieren zwischen Klassenstufen nachweisen. In einer vergleichenden Untersuchung von Mathematikaufgaben der Klassenstufen 4, 7 und 10 aus dem U.S.-Bundesstaat Kansas zeigten Shaftel et al. (2006) zwar steigende sprachliche Anforderungen über die Klassenstufen hinweg. Statistisch bedeutsame Effekte der nicht-mathematischen sprachlichen Merkmale auf die Aufgabenschwierigkeit konnten jedoch ausschließlich in der vierten Klassenstufe festgestellt werden. Die Autoren konnten außerdem keine differenziellen Effekte der sprachlichen Anforderungen für ELLs finden, was darauf hinweist, dass sowohl ELLs als auch nicht-ELLs in unteren

Klassenstufen stärker durch sprachliche Anforderungen von Testaufgaben beeinflusst werden als in höheren Klassenstufen.

Die in der Teilstudie 4 der vorliegenden Arbeit erzielten Befunde zu der Frage, ob die sprachlichen Anforderungen der Aufgaben in der dritten Klassenstufe einen anderen Effekt auf die differenzielle Validität von Mathematiktests für Schülerinnen und Schüler mit nicht-deutscher Familiensprache ausüben als in der vierten Klassenstufe, decken sich mit den Befunden von Shaftel et al. (2006). In dieser Teilstudie zeigten sich sowohl für Schülerinnen und Schüler mit nicht-deutscher Familiensprache als auch für Schülerinnen und Schüler mit deutscher Familiensprache für die dritte Klassenstufe stärkere Zusammenhänge der Lösungswahrscheinlichkeit von Mathematikaufgaben mit den sprachlichen Anforderungen der Aufgaben als für die vierte Klassenstufe. Dieser Befund deutet darauf hin, dass jüngere Grundschulkinder unabhängig von ihrer Familiensprache größere Probleme hatten, bildungssprachliche Anforderungen von Mathematikaufgaben zu bewältigen als ältere Grundschulkinder.

Diese Probleme könnten auf die bei jüngeren Grundschulkindern im Durchschnitt geringere Lesefähigkeit zurückzuführen sein. Beispielsweise konnten Bremerich-Vos und Böhme (2009) in der Normierungsstudie für die Aufgaben des IQB-Ländervergleichs Primarstufe eine um d = 0.60 geringere Lesefähigkeit bei Schülerinnen und Schülern der dritten Klasse im Vergleich zu Schülerinnen und Schülern der vierten Klasse feststellen. Aus Teilstudie 1 ergab sich für die Daten des IQB-Ländervergleichs Primarstufe 2011, dass Schülerinnen und Schüler mit nicht-deutscher Familiensprache in der vierten Klasse eine um d = 0.32 geringere Lesefähigkeit aufwiesen als Schülerinnen und Schüler mit deutscher Familiensprache. Der mittlere Unterschied in der Lesefähigkeit zwischen Schülerinnen und Schülern verschiedener Klassenstufen fiel für die betrachteten Daten somit größer aus als der mittlere Unterschied zwischen Schülerinnen und Schülern mit nicht-deutscher Familiensprache und Schülerinnen und Schülern mit deutscher Familiensprache in der vierten Klasse. In Teilstudie 3 konnte außerdem gezeigt werden, dass die Wirksamkeit einer sprachlichen Vereinfachung von Testaufgaben von der Lesefähigkeit der Schülerinnen und Schüler abhängt. Diese Befunde könnten dahingehend interpretiert werden, dass die Lesefähigkeit der Schülerinnen und Schüler an der Entstehung einer differenziellen Validität der Aufgaben beteiligt ist und das der Einfluss der sprachlichen Anforderungen auf die differenzielle Validität auch von der Lesefähigkeit abhängt (vgl. Grießhaber, 2011). Das Ausmaß dieser Abhängigkeit sollte in weiteren Studien genauer untersucht werden.

4.1.3 Befunde zum Einfluss einzelner bildungssprachlicher Merkmale der Aufgaben auf differenzielle Validität

Bezüglich des Einflusses einzelner bildungssprachlicher Merkmale konnte die vorliegende Arbeit in den Teilstudien 1 und 3 spezifische Effekte einzelner deskriptiver (Textlänge), lexikalischer (allgemeiner und mathematikspezifischer bildungssprachlicher Wortschatz) und grammatischer (Anzahl der Nominal- und Präpositionalphrasen) Merkmale nachweisen. Allerdings zeigte sich in Teilstudie 3, dass der überwiegende Teil differenziellen Itemfunktionierens durch mehrere Merkmale gemeinsam aufgeklärt wird. Auch in Teilstudie 1 zeigten sich für mehrere bildungssprachliche Merkmale ähnliche Zusammenhänge mit differenziellem Itemfunktionieren.

Bisherige nicht-experimentelle Arbeiten zu bildungssprachlicher Komplexität unterscheiden sich in der Operationalisierung bildungssprachlicher Anforderungen von Testaufgaben. Die bildungssprachlichen Anforderungen der Aufgaben wurden vergleichsweise selten direkt mit globalen Expertenurteilen erfasst (Martiniello, 2009). Diese Art der Erfassung hat den Vorteil, dass in den Analysen nur eine Variable für die bildungssprachlichen Anforderungen berücksichtigt werden muss. Als Nachteil dieses Verfahrens kann die undifferenzierte Erfassung bildungssprachlicher Komplexität gesehen werden. In den meisten Arbeiten wird die Häufigkeit bestimmter bildungssprachlicher Merkmale ausgezählt, wodurch die sprachlichen Anforderungen differenzierter erfasst werden können (Abedi et al., 1997; Bailey et al., 2007; Shaftel et al., 2006; Spanos, Rhodes, Dale & Crandall, 1988; Wolf & Leon, 2009). Diese Merkmale werden in einigen Arbeiten zu übergeordneten Maßen für sprachliche Komplexität zusammengefasst (Abedi et al., 1997; Wolf & Leon, 2009).

In bisherigen Arbeiten zur sprachlichen Komplexität von Aufgaben wurde jedoch nicht untersucht, ob die Zusammenfassung zu übergeordneten Maßen aus statistischer Sicht gerechtfertigt ist. Für den Bereich der Sprachproduktion von Vorschulkindern konnten Leseman et al. (2007) sowie Scheele et al. (2012) zeigen, dass sich die bildungssprachlichen Merkmale stark überlappen und dass die sprachliche Komplexität faktorenanalytisch auf einen Hauptfaktor reduziert werden kann. Die vorliegende Arbeit konnte in Übereinstimmung mit diesem Befund zeigen, dass sich die einzelnen bildungssprachlichen Merkmale für die verwendeten Testaufgaben im Fach Mathematik stark überlappen und empirisch kaum voneinander trennbar sind. Außerdem konnte aus den einzelnen Merkmalen faktorenanalytisch ein Globalfaktor bildungssprachlicher Komplexität extrahiert werden, der sich differenziell nach Klassenstufe auf die Lösungswahrscheinlichkeit von Mathematikaufgaben auswirkt. Diese Befunde deuten darauf hin, dass die bildungssprachlichen Anforderungen von Testaufgaben in der Primarstufe als ein homogenes Konstrukt konzeptionalisiert wer-

den können. Diese Interpretation wird auch durch die für die verschiedenen Itemstichproben übereinstimmend hohen Interkorrelationen zwischen den bildungssprachlichen Merkmalen gestützt, die in den Teilstudien 1 (für den Bereich Lesen) sowie 2 und 4 (für den Bereich Mathematik) gefunden wurden.

Für zukünftige Forschungsvorhaben, in denen der globale Einfluss sprachlicher Anforderungen von Testaufgaben auf die Lösungswahrscheinlichkeit dieser Aufgaben geprüft werden soll, könnte es daher günstig sein, bildungssprachliche Anforderungen von Testaufgaben der Primarstufe mit einer globalen Einschätzung zu bestimmen. Diese globale Einschätzung wäre ökonomischer als die Erfassung von Einzelkriterien, die später zusammengefasst werden (vgl. Martiniello, 2009).

Dieses Vorgehen könnte allerdings auch dazu führen, dass mögliche Klassenstufenunterschiede in den sprachlichen Anforderungen nur ungenau abgebildet werden können, wenn für mehrere Klassenstufen dasselbe Raster zur globalen Einschätzung verwendet wird. Unterschiede in der Bedeutsamkeit einzelner Merkmale zwischen Klassenstufen könnten dann nicht mehr festgestellt werden. Die Studien von Deane et al. (2006) sowie Graesser et al. (2011) deuten darauf hin, dass die Häufigkeit der einzelnen Merkmale sprachlicher Anforderungen englischsprachiger Lesetexte über die Schulzeit hinweg unterschiedlich stark zunimmt. Dadurch veränderte sich das Verhältnis der Häufigkeit verschiedener Merkmale über die Schulzeit. Daher sollte ein Globalscore sprachlicher Komplexität für verschiedene Klassenstufen unterschiedlich zusammengesetzt sein. Dies würde bedeuten, dass die Merkmale je nach Klassenstufe mit unterschiedlicher Gewichtung in die Bildung eines Bildungssprachmaßes eingehen sollten, wobei die Bedeutung einzelner Merkmale in spezifischen Klassenstufen zunächst theoretisch und empirisch bestimmt werden müsste.

4.2 Grenzen der vorliegenden Arbeit

Die Grenzen der vorliegenden Arbeit beziehen sich auf die näherungsweise Operationalisierung der Lesefähigkeiten im Deutschen der Schülerinnen und Schüler über den Indikator „nicht-deutsche Familiensprache“, auf den in der vorliegenden Arbeit verwendeten Aufgabenpool und auf die Operationalisierung von differenzieller Validität mittels differenziellem Itemfunktionieren.

In der Literatur wird angenommen, dass bildungssprachliche Merkmale von Mathematikaufgaben besonders diejenigen Zweitsprachlernenden benachteiligen, die über mittlere bis eher geringe sprachliche Fähigkeiten in der Testsprache verfügen (z. B. Grießhaber, 2011). Da die Aufgaben in schriftsprachlicher Form vorliegen, wäre die Lesefähigkeit ein möglicher Indikator für die benötigten sprachlichen Fähigkeiten in der Testsprache Deutsch. In der

vorliegenden Arbeit wurde die Lesefähigkeit im Deutschen der Schülerinnen und Schüler nur in der Teilstudie 3 direkt operationalisiert. In den Teilstudien 2 und 4 wurde stattdessen die Familiensprache der Schülerinnen und Schüler als Gruppierungsvariable verwendet. In der vorliegenden Arbeit war die direkte Einbeziehung der Lesefähigkeit als Prädiktor für die Teilstudien 2 und 3 nicht möglich, da im Datensatz der Teilstudie 2 kein Indikator für die Lesefähigkeit vorhanden war und für Teilstudie 3 nur für die Schülerinnen und Schüler der vierten Klasse ausreichende Daten für die Lesefähigkeit vorlagen.

In den für die vorliegende Arbeit verwendeten Stichproben verfügen Schülerinnen und Schüler mit nicht-deutscher Familiensprache im Mittel über geringere sprachliche Kompetenzen in den Bereichen Lesen und Zuhören als Schülerinnen und Schüler mit deutscher Familiensprache (IQB-Ländervergleich 2011: Böhme & Weirich, 2012; VERA: Kuhl, Harych & Vogt, 2011, siehe auch Abschnitt 4.1.2). Daher kann die Familiensprache als ein Indikator für geringe sprachliche Kompetenzen in diesen Bereichen dienen. Ferner konnte mit Daten aus PISA-E 2000 gezeigt werden, dass der Effekt der Familiensprache auf die Lesekompetenz auch nach Kontrolle von sozialen (wie z. B. dem SES der Familie) und motivationalen Merkmalen der Schülerinnen und Schüler noch statistisch bedeutsam ist (Baumert, Watermann & Schümer, 2003; vgl. auch A. G. Müller & Stanat, 2006). Daher kann eine nicht-deutsche Familiensprache der Schülerinnen und Schüler als Indikator für geringere Lesefähigkeiten im Vergleich zu monolingual deutschsprachig aufwachsenden Schülerinnen und Schülern gesehen werden.

Eine direktere Operationalisierung sprachlicher Fähigkeiten wäre für zukünftige Forschungsvorhaben allerdings dennoch wünschenswert, da aus dem Indikator Familiensprache nur mit einer gewissen Unsicherheit auf diese Fähigkeiten geschlossen werden kann. So zeigen beispielsweise die Befunde der IGLU-Studie, dass Kinder mit nicht-deutscher Familiensprache zwar im Mittel niedrigere Leistungen im Lesen erzielen, sich die Verteilungen der Lesewerte für Kinder mit deutscher Familiensprache und Kinder mit nicht-deutscher Familiensprache aber deutlich überlappen (Schwippert et al., 2012). Dies deutet darauf hin, dass die Lesefähigkeit innerhalb der Gruppe der Schülerinnen und Schüler mit nicht-deutscher Familiensprache heterogen ausgeprägt ist, sodass auch Schülerinnen und Schüler mit nicht-deutscher Familiensprache hohe Lesefähigkeiten erreichen können. Bei Schülerinnen und Schülern mit hoher Lesefähigkeit ist jedoch nicht davon auszugehen, dass die sprachlichen Anforderungen der Mathematikaufgaben benachteiligend wirken. Diese innerhalb der Gruppe der Schülerinnen und Schüler mit nicht-deutscher Familiensprache bestehende Heterogenität bezüglich der Lesefähigkeit in der Testsprache Deutsch könnte dazu führen, dass eine eventuell bestehende Benachteiligung einer Subgruppe von Schülerinnen und Schülern

mit geringen Lesefähigkeiten durch sprachliche Anforderungen von Testaufgaben nicht entdeckt wird (vgl. auch Abschnitte 4.1.2 und 4.3).

In Teilstudie 1 konnte gezeigt werden, dass auch die Messung der Lesefähigkeit von den bildungssprachlichen Anforderungen der Aufgaben differenziell beeinflusst wird. Die bildungssprachlichen Merkmale der Leseaufgaben stellten Schülerinnen und Schüler mit nichtdeutscher Familiensprache vor größere Schwierigkeiten als Schülerinnen und Schüler mit deutscher Familiensprache. Daher wäre eine direkte Messung bildungssprachlicher Fähigkeiten der Schülerinnen und Schüler mit geeigneten Testverfahren die beste Operationalisierung der theoretisch angenommenen besonders schulrelevanten Sprachfähigkeiten. Entsprechend konnten Prediger et al. (2013) deutliche Leistungsunterschiede in einem Mathematiktest zwischen zwei Gruppen von Schülerinnen und Schülern identifizieren, die sich in ihren bildungssprachlichen Fähigkeiten unterschieden. Für Gruppierungen, die anhand eines nicht speziell auf Bildungssprache ausgerichteten Lesetests oder anhand des Zuwanderungshintergrunds der Schülerinnen und Schüler gebildet wurden, zeigten sich hingegen geringere Leistungsunterschiede. Es wäre zu prüfen, inwieweit sich auch itemspezifische Leistungsunterschiede zwischen diesen verschiedenen Gruppen zeigen, die auf spezifische Benachteiligungen hinweisen könnten.

In den Teilstudien 1, 2 und 4 der vorliegenden Arbeit wurden bereits existierende Aufgaben aus den Aufgabenpools von VERA-3 sowie dem IQB-Ländervergleich 2011 verwendet. Ein Teil der Aufgaben aus dem IQB-Ländervergleich 2011 wurde im Rahmen der Entwicklung und Erprobung dieser Aufgaben bereits vor Beginn der vorliegenden Studie auf ihre Fairness geprüft. Zu diesem Zweck wurden auf Basis der Daten einer Pilotierungsstudie DIF-Analysen für Kinder mit nicht-deutscher Familiensprache durchgeführt. Vor allem die Aufgaben mit auffälligen DIF-Werten wurden dahingehend überarbeitet, dass offensichtliche sprachliche Hürden für Kinder mit nicht-deutscher Familiensprache minimiert wurden. Die sprachlichen Hürden wurden für diese Überarbeitung allerdings nicht nach dem Konzept der Bildungssprache definiert. Daher war zu erwarten, dass die Effekte der bildungssprachlichen Anforderungen für die Aufgaben des Ländervergleichs 2011 geringer ausfallen als es bei nicht überarbeiteten, sprachlich komplexeren Aufgaben der Fall wäre.

Im Rahmen der vorliegenden Arbeit konnten trotz dieser Überarbeitungen im Entwicklungsprozess heterogene bildungssprachliche Anforderungen der Aufgaben des IQB-Ländervergleichs 2011 gezeigt werden. Ferner ermittelten Studien aus dem amerikanischen Sprachraum auch für Aufgaben aus großen Schulleistungsstudien substanzielle, durch die sprachlichen Anforderungen der Testaufgaben bedingte, Nachteile für English Language

Learners (Abedi & Lord, 2001; Abedi et al., 1997; Noble, Rosebery, Suarez, Warren & O'Connor, 2014). Ein Vorteil der Verwendung realer Testaufgaben in der vorliegenden Arbeit ist, dass die so ermittelten sprachlichen Anforderungen sowie deren Zusammenhänge mit differenziellen Leistungsnachteilen für Kinder mit nicht-deutscher Familiensprache ökologisch valide sind, da sie sich auf alle tatsächlich eingesetzten Aufgaben einer großen nationalen Schulleistungsstudie beziehen. Die festgestellte weitgehende Abwesenheit von differenzieller Validität der Aufgaben des IQB-Ländervergleichs 2011 kann daher dahingehend interpretiert werden, dass die vorgenommenen Gruppenvergleiche von Kindern mit deutscher Familiensprache und Kindern mit nicht-deutscher Familiensprache valide und nicht in substanziellem Maß auf eine differenzielle Validität des Tests für Kinder mit nicht-deutscher Familiensprache zurückzuführen sind.

Die vorliegende Arbeit prüft die differenzielle Validität der Tests auf der Ebene von Einzelitems. Das gewählte Verfahren des differenziellen Itemfunktionieren erlaubt keine Prüfung der Fairness auf der Ebene des Gesamttests, da der DIF per Definition im Mittel 0 beträgt. Der Grund für dieses Vorgehen war, dass sich die Hypothesen explizit auf die sprachliche Komplexität der Testitems beziehen und somit itemspezifische Schwierigkeitsunterschiede zwischen Gruppen im Fokus der Arbeit standen. Eine Fairnessprüfung auf der Ebene des Gesamttests könnte beispielsweise im Rahmen von *Score Equity Analysen* (vgl. Dorans, 2004; Huggins, 2014; Huggins & Elbaum, 2013) vorgenommen werden. Dabei wird geprüft, inwiefern die Anbindung an eine Normpopulation für verschiedene Gruppen zu unterschiedlichen Ergebnissen führt. Eine genauere Beschreibung des Vorgehens findet sich in Abschnitt 2.2.1 der vorliegenden Arbeit. Im Rahmen der vorliegenden Arbeit wäre dies nur für die Daten der Teilstudie 2 möglich gewesen, da die anderen, auf dem IQB-Ländervergleich 2011 beruhenden Teilstudien die Normpopulation umfassen. Mit Score Equity Analysen ließe sich prüfen, ob der Test im Ganzen Schülerinnen und Schüler mit nicht-deutscher Familiensprache substanziell benachteiligt, beispielsweise weil alle Aufgaben die sprachlichen Fähigkeiten dieser Gruppe übersteigen. In der vorliegenden Arbeit konnte allerdings gezeigt werden, dass sich die Aufgaben in ihren bildungssprachlichen Anforderungen deutlich unterscheiden und dass die sprachlichen Anforderungen insgesamt eher gering ausgeprägt sind. Daher erscheint die Annahme, dass alle Aufgaben zu hohe sprachliche Anforderungen für Kinder mit nicht-deutscher Familiensprache aufweisen, aufgrund der empirischen Ergebnisse zu Anzahl und Verteilung der bildungssprachlichem Merkmale im Aufgabenmaterial nicht plausibel.

In der vorliegenden Arbeit wurden Unterschiede in den Effekten sprachlicher Komplexität zwischen Kindern mit nicht-deutscher Familiensprache und Kindern mit deutscher Famili-

ensprache geprüft. Diese Gruppierung wurde gewählt, um die Fairness der Testverfahren für eine Gruppe von Testteilnehmern zu prüfen, die in der Berichtlegung solcher Studien getrennt ausgewiesen wird (vgl. Zumbo, 2007) und für die in bisherigen Schulleistungsstudien Disparitäten auftraten (Bos, Tarelli, et al., 2012; Bos, Wendt, et al., 2012; Pant et al., 2013; Prenzel et al., 2013; Stanat et al., 2012). Bei Fairnessprüfungen wird implizit angenommen, dass sich die Personen bezüglich für den Test relevanter Merkmale, wie beispielsweise Antwortprozesse bei der Bearbeitung der Aufgaben, innerhalb der Gruppen ähnlicher sind als zwischen Gruppen (Camilli, 2006). Es gibt allerdings Hinweise darauf, dass die Gruppe der Schülerinnen und Schüler mit nicht-deutscher Familiensprache nicht homogen ist, sondern dass sich die Leistungen in Mathematik und in Lesen beispielsweise deutlich zwischen den verschiednenen Herkunftsgruppen unterscheiden (Haag et al., 2012; Stanat et al., 2010). Für die vorliegende Arbeit wurde allerdings lediglich die Gesamtgruppe aller Kinder mit nicht-deutscher Familiensprache betrachtet. Dies lässt sich mit der Datenlage begründen. Für den Datensatz der Teilstudie 2 sowie für den Teildatensatz der dritten Klasse der Teilstudie 4 lagen keine Angaben über die Herkunftsgruppe der Schülerinnen und Schüler vor, wohingegen für den Teildatensatz der vierten Klasse der Teilstudie 4 aufgrund der hohen Itemanzahl von über 200 Items die Anzahl der Itemantworten für die einzelnen Herkunftsgruppen zu klein gewesen wäre, um sie im Rahmen von DIF-Analysen reliabel getrennt auswerten zu können. Allerdings wurde der SES der Schülerinnen und Schüler als ein wichtiges Merkmal der Heterogenität in die Analysen mit einbezogen, soweit dies aufgrund der Datenlage möglich war.

In heterogenen Gruppen kann es zu Problemen bei der Feststellung und der Erklärung von DIF-Effekten kommen, da sich moderierende Variablen, wie z. B. die Lesefähigkeit, zwischen den einzelnen Subgruppen unterscheiden können (Camilli, 2013; Ercikan, Roth, Simon, Sandilands & Lyons-Thomas, 2014; Oliveri, Ercikan & Zumbo, 2014; Sireci & Rios, 2013). Als Abhilfe wurden verschiedene Möglichkeiten vorgeschlagen, homogenere Gruppen für DIF-Analysen zu identifizieren. Hierzu bietet sich beispielsweise die Zerlegung der Gesamtgruppe nach mehr als einer DIF-Variable an (z. B. Ercikan et al., 2014; Oliveri, Ercikan & Zumbo, 2014), wobei neben individuellen Merkmalen wie beispielsweise Geschlecht oder Familiensprache auch institutionelle Faktoren wie Schulform oder curriculare Bedingungen berücksichtigt werden sollten (vgl. McElvany & Schwabe, 2013; Schwabe & Gebauer, 2013; Schwabe, McElvany & Trendtel, 2015, im Druck). Alternativ können Gruppen von Personen gebildet werden, die sich anhand ihrer Antwortprozesse unterscheiden. Die Analyse von Antwortprozessen kann mit Hilfe latenter Klassenanalysen auch für nicht im Vorhinein identifizierte Gruppen geschehen (z. B. Oliveri, Ercikan, Zumbo & Lawless, 2014). In einem

weiteren Schritt kann dann geprüft werden, welche Variablen die Zugehörigkeit zu den latenten Klassen vorhersagen. Mit diesem Ansatz konnten Sandilands, Oliveri, Zumbo und Ercikan (2013) beispielsweise zeigen, dass DIF-Werte zwischen Ländern auf Aspekte der Unterrichtsgestaltung zurückgeführt werden können. Die Verwendung von latenten Klassenanalysen bietet vielversprechende Möglichkeiten, in zukünftigen Studien differenziertere Einblicke zu bekommen, welche Schülergruppen durch sprachliche Merkmale von Testaufgaben benachteiligt werden.

4.3 Implikationen für die Praxis

Aus der vorliegenden Arbeit ergeben sich verschiedene Implikationen für die Testkonstruktion, für die Interpretation der Befunde des IQB-Ländervergleichs in der Primarstufe und für eine homogenere Definition von Zweitsprachlernenden für die Ergebnisberichte von Schulleistungsstudien.

Zum einen lässt sich für die Testkonstruktion ableiten, dass sich die im IQB verwendeten Testkonstruktionsprozesse bewährt haben. Es konnten Tests bereitgestellt werden, welche die Kompetenzen der für die Berichtlegung relevanten Schülergruppen fair messen. Dies zeigt sich an dem Befund, dass Schülerinnen und Schüler mit nicht-deutscher Familiensprache sowie Schülerinnen und Schüler aus Familien mit niedrigem SES durch die sprachlichen Merkmale der Mathematikaufgaben nur in vernachlässigbarem Ausmaß benachteiligt werden. Ein Grund hierfür ist, dass einige der von Camilli (2013) empfohlenen Schritte zur Fairnesssicherung bereits im Prozess der Testkonstruktion am IQB berücksichtigt wurden. Bei diesen berücksichtigten Schritten handelt es sich um die Anbindung der Testkonstruktion an ein Curriculum bzw. an die nationalen Bildungsstandards, umfassende Analysen des Aufgabeninhalts durch fachdidaktische Kooperationspartner, groß angelegte Erprobungsstudien und, im Fall der Aufgaben für den IQB-Ländervergleich 2011, DIF-Analysen für Kinder mit nicht-deutscher Familiensprache. Außerdem wurde darauf geachtet, dass die in den Aufgaben verwendeten Situationen an die Lebenswelt der Schülerinnen und Schüler angepasst sind. Dieser Punkt greift die Kritik von Solano-Flores und Trumbull (2003) an Testkonstruktionsprozessen auf, die darauf hinweisen, dass kulturelle und sprachliche Unterschiede in der Schülerschaft schon im Konstruktionsprozess der Tests systematisch berücksichtigt werden sollten (vgl. auch Lane & Leventhal, 2015). Die genannten Maßnahmen, um möglichst faire Items zu entwickeln, sollten daher auch in zukünftigen Testentwicklungsprozessen umgesetzt werden.

Die vorliegende Arbeit konnte außerdem zeigen, dass die im Rahmen des IQB-Ländervergleichs 2011 und der VERA-3 Erhebungen festgestellten sozialen und zuwande-

rungsbezogenen Disparitäten ein valides Abbild der Leistungsunterschiede darstellen. Insbesondere fanden sich keine Belege für die populäre Ansicht, dass Kinder mit nicht-deutscher Familiensprache durch die Tests substanziell benachteiligt werden und dass die Testergebnisse für diese Schülergruppe daher nicht aussagekräftig sind. Dieser Befund impliziert, dass die Ergebnisse der genannten Schulleistungsstudien als Ausgangspunkt für Schul- und Unterrichtsentwicklung genutzt werden können. Die Kehrseite dieses Befundes ist, dass sich die Leistungsunterschiede zwischen Schülerinnen und Schülern mit deutscher Familiensprache und Schülerinnen und Schülern mit nicht-deutscher Familiensprache nicht durch die verwendeten Testaufgaben erklären lassen. Daher bleibt weiterhin offen, ob bzw. wie Kompetenzunterschiede durch die Familiensprache bedingt werden. Eine Möglichkeit wäre, dass die sprachliche Fähigkeit nicht primär in der Testsituation zu Nachteilen führt, sondern schon im Unterrichtsgeschehen. Dies könnte dann der Fall sein, wenn Kinder mit nicht-deutscher Familiensprache Probleme haben, dem Unterricht zu folgen und die unterrichteten mathematischen Konzepte zu erwerben (vgl. Duarte et al., 2011; Grießhaber, 2011).

Ein Einfluss von sprachlichen Fähigkeiten in der Unterrichtssituation wird durch einige Studien nahegelegt, die sprachassoziierte Unterschiede in mathematischen Fähigkeiten feststellen konnten. Sprachliche Fähigkeiten scheinen sowohl bei Erstsprachlernern als auch bei Zweitsprachlernenden schon vor dem Eintritt in die Grundschule einen Einfluss auf die Entwicklung mathematischer Fähigkeiten auszuüben. In einer querschnittlichen Studie mit 130 Kindergartenkindern aus den Niederlanden konnten Kleemans, Segers und Verhoeven (2011) beispielsweise zeigen, dass frühe sprachliche Fähigkeiten—phonologische Bewusstheit und grammatische Fähigkeiten—und frühe mathematische Fähigkeiten—logische Operationen und Zählen—bei Erstsprachlernenden und bei Zweitsprachlernenden vor dem Schuleintritt ungefähr gleich hoch miteinander zusammenhingen. Die Leistungsunterschiede zwischen Erst- und Zweitsprachlernenden lassen sich teilweise durch Unterschiede zwischen den Gruppen in der Verfügbarkeit lernförderlicher familiärer Aktivitäten, wie beispielsweise die Förderung von Lesen, Buchstabieren, Zählen und Benennen von Formen, erklären (Anders et al., 2012). Hierbei zeigten sprachliche Aktivitäten sogar noch stärkere Zusammenhänge mit den mathematischen Vorläuferfähigkeiten als mathematische Aktivitäten, was wiederum auf die wichtige Rolle sprachlicher Fähigkeiten beim Erwerb mathematischer Fähigkeiten hindeutet.

Auch in der Grundschulzeit zeigten sich sprachassoziierte Unterschiede in mathematischen Fähigkeiten (Heinze, Herwartz-Emden, Braun & Reiss, 2011; Heinze, Herwartz-Emden & Reiss, 2007; Hickendorff, 2013; Reardon & Galindo, 2009; Vukovic & Lesaux, 2013). Das

Muster der Beziehungen zwischen sprachlichen und mathematischen Fähigkeiten war zu Beginn der Grundschulzeit für Erst- und Zweitsprachlernende zwar ähnlich, die Beziehungen waren aber für Zweitsprachlernende tendenziell stärker ausgeprägt (Vukovic & Lesaux, 2013). Vergleicht man diesen Befund mit für die Vorschule festgestellten gleich stark ausgeprägten Beziehungen zwischen mathematischen und sprachlichen Fähigkeiten für Erst- und Zweitsprachlernenden, so könnten die stärkeren Zusammenhänge für Zweitsprachlernende in der Grundschule darauf hindeuten, dass diese Schülerinnen und Schüler bereits in der Grundschule nicht optimal vom Mathematikunterricht profitieren und daher tendenziell niedrigere Leistungszuwächse verzeichnen.

Ferner sollten für die Berichtlegung von Schulleistungsstudien innerhalb der Gruppe der Kinder mit nicht-deutscher Familiensprache stärker diejenigen Kinder in den Blick genommen werden, die nicht über ausreichende Sprachkenntnisse in der Instruktionssprache verfügen. Diese Gruppe wäre im Vergleich mit der Gruppe aller Kindern mit nicht-deutscher Familiensprache homogener im Hinblick auf ihre sprachlichen Voraussetzungen in der Instruktionssprache. Dadurch könnten eventuelle Leistungsrückstände dieser Gruppe möglicherweise eindeutiger auf geringere Deutschkenntnisse zurückgeführt werden und dadurch besser interpretierbar sein.

Die in den USA verwendete Definition von Zweitsprachlernenden mit sprachlichem Förderbedarf (English Language Learners, ELL) bezieht den Sprachstand der Schülerinnen und Schüler explizit mit ein. Diese Definition ist für die vorliegende Arbeit relevant, da ein Großteil der Forschung zur möglichen Benachteiligung von Zweitsprachlernenden durch sprachlich komplexe Textaufgaben aus den USA stammt und mit English Language Learners (ELLs) durchgeführt wurde. Unter ELLs werden Schülerinnen und Schüler mit Zuwanderungshintergrund verstanden, deren Englischkenntnisse zu gering sind, um dem normalen Regelunterricht ohne zusätzliche Hilfe folgen zu können (Abedi, 2008a). ELLs werden—allerdings mit beträchtlichen Unterschieden zwischen den Staaten—durch standardisierte Tests der Englischkenntnisse festgestellt, wobei auch diese nicht immer ausreichend auf die in der Schule benötigten Englischkenntnisse zugeschnitten sind (Abedi, 2008b; Bailey & Huang, 2011; Butler et al., 2007). Im Gegensatz zu den demografischen Merkmalen Zuwanderungsgeneration, Herkunftsgruppe und Familiensprache handelt es sich bei ELLs um eine Subgruppe von Schülerinnen und Schülern mit Zuwanderungshintergrund, die einen zusätzlichen zweitsprachlichen Förderbedarf haben. Ziel des Schulsystems ist es, dass diese Schülerinnen und Schüler durch speziell auf ihre Bedürfnisse zugeschnittenen Förderunterricht ausreichende Sprachkenntnisse erwerben, um am regulären Unterricht teilnehmen zu können. Die Schätzungen der dafür im Durchschnitt benötigten Zeit belaufen sich auf ca. fünf bis

neun Jahre (Hakuta, Butler & Witt, 2000; Slama, 2012, 2014). Eine vergleichbare Diagnostik inklusive dem dazugehörigen Fördersystem gibt es in Deutschland nicht, möglicherweise auch deshalb, weil die Zahl von Zuwanderern der ersten Generation im Schulsystem bislang vergleichsweise niedrig ist. Es wäre wünschenswert, eine solche Definition auch in Deutschland einzuführen, da damit eine homogenere Gruppe von Zweitsprachlernenden mit potenziellen sprachlichen Problemen identifiziert werden könnte (vgl. auch Prediger, Wilhelm, Büchter, Gürsoy & Benholz, 2015).

4.4 Implikationen für die Forschung und Ausblick auf zukünftige Forschungsfragen

Anknüpfend an diese Arbeit lassen sich Forschungsfragen bezüglich des Verhältnisses von mathematischen und sprachlichen schwierigkeitsgenerierenden Merkmalen von Mathematikaufgaben, der Operationalisierung bildungssprachlicher Merkmale sowie des Verhältnisses der einzelnen bildungssprachlichen Merkmale zueinander und zur Messung bildungssprachlicher Schülerfähigkeiten ableiten.

4.4.1 Verhältnis von sprachlichen und mathematischen Anforderungen der Testaufgaben

In zukünftigen Forschungsarbeiten sollte kritisch hinterfragt werden, inwieweit bildungssprachliche Merkmale von Testaufgaben als konstruktirrelevant zu betrachten sind. Im Sinne von Messick (1989) können sprachliche Anforderungen von Testaufgaben nur zu Validitätsproblemen führen, wenn sie für das zu messende Konstrukt irrelevant sind. Vor allem die Arbeitsgruppe um Jamal Abedi argumentiert, dass bestimmte sprachliche Merkmale der Aufgaben als konstruktirrelevant betrachtet werden sollten (Abedi, 2002, 2009; Abedi & Gándara, 2006; Abedi & Lord, 2001; Abedi et al., 1997).

Allerdings sind sprachliche und mathematische Fähigkeiten bereits aus der normativen Perspektive der nationalen Bildungsstandards miteinander verbunden (vgl. Walker, Zhang & Surber, 2008). Die in den Bildungsstandards beschriebenen mathematischen Kompetenzen, die Schülerinnen und Schüler am Ende der Primarstufe erworben haben sollen, stellen eine „Grundlage für das Mathematiklernen in den weiterführenden Schulen und für die lebenslange Auseinandersetzung mit mathematischen Anforderungen des täglichen Lebens“ (KMK, 2004, S. 6) dar. Diese Grundlage umfasst neben dem Beherrschen arithmetischer Grundoperationen beispielsweise auch den Umgang mit in alltäglichen Situationen auftretenden angewandten mathematischen Problemen. Entsprechende Aufgaben werden für gewöhnlich sprachlich dargeboten, sodass sprachliche Fähigkeiten zum Verstehen und Lösen dieser Probleme sowie zur Interpretation der Lösung unabdingbar sind. Sprachliche Kompetenzen

werden außerdem beispielsweise in den Standards zu den Bereichen „mathematisch kommunizieren“ und „mathematisch argumentieren“ explizit gefordert (vgl. Linneweber-Lammerskitten, 2013). Die am IQB entwickelten Mathematiktests sollen die nationalen Bildungsstandards messbar machen und beinhalten daher neben prozeduralen Routineaufgaben auch an sprachlich vermittelten mathematikhaltigen Situationen orientierte Sachaufgaben. Eine Messung der in den Bildungsstandards beschriebenen mathematischen Kompetenzen, die ausschließlich mit textfreien Mathematikaufgaben durchgeführt wird, würde daher das Konstrukt nicht in vollem Umfang operationalisieren können. In diesem Fall wäre die Validität des Tests eingeschränkt, da wichtige Teile des Konstrukts im Test unberücksichtigt blieben (construct underrepresentation, Messick, 1989). Zukünftige Forschung sollte sich daher aus fachdidaktischer Sicht mit dem notwendigen Ausmaß sprachlicher Einbettung von Mathematikaufgaben auseinandersetzen (vgl. auch Avenia-Tapper & Llosa, 2015).

Außerdem sollte geprüft werden, inwiefern sprachliche und mathematische Merkmale der Aufgaben die Aufgabenschwierigkeit beeinflussen und inwiefern sich die Gewichtung dieser Merkmale für Schülerinnen und Schüler deutscher Familiensprache und für Schülerinnen und Schüler nicht-deutscher Familiensprache unterscheidet. Die vorliegende Arbeit konzentrierte sich auf die sprachlichen Aspekte der Aufgaben. Daher wurden keine nichtsprachlichen mathematischen Schwierigkeiten der Mathematikaufgaben berücksichtigt oder systematisch variiert. Es erscheint zum einen möglich, dass konzeptuell schwierigere Aufgaben auch mit höheren sprachlichen Anforderungen einhergehen (vgl. Mullis, Martin & Foy, 2013) und zum anderen, dass Kinder mit nicht-deutscher Familiensprache mit konzeptuell schwierigeren Aufgaben größere Probleme haben könnten (Duarte et al., 2011; Heinze et al., 2011; Pimperton & Nation, 2010; Prediger et al., 2013; Prediger et al., 2015; Ufer, Reiss & Mehringer, 2013). Beispielsweise fanden Wolf und Leon (2009) einen Moderatoreffekt der Aufgabenschwierigkeit auf die Beziehung zwischen sprachlichen Merkmalen und differenziellen Schwierigkeiten von ELLs. Sie konnten feststellen, dass sprachliche Variablen stärker mit DIF zuungunsten von ELLs auf einfachen Items als auf schwierigen Items korrelieren. Dies deutet darauf hin, dass eine falsche Lösung bei mathematisch einfachen Items bei Zweitsprachlernenden eher auf sprachliche Merkmale der Aufgabe zurückzuführen sein könnte, wohingegen bei mathematisch schwierigen Aufgaben sowohl Erst- als auch Zweitsprachlernende eher an dem mathematischen Gehalt der Aufgabe scheitern. Im Rahmen generalisierter gemischter Modelle könnte ein solcher Effekt als Dreifachinteraktion zwischen den Aufgabenmerkmalen „sprachliche Komplexität“ und „mathematische Komplexität“ mit dem Personenmerkmal „Familiensprache“ modelliert werden.

Ferner sollte erforscht werden, wie sich die in der Mathematik verwendete Bildungssprache von der Bildungssprache anderer Fächer abgrenzen lässt. Verschiedene Autoren haben fachspezifische bildungssprachliche Merkmale für Mathematik beschrieben (z. B. Rösch & Paetsch, 2011; Schleppegrell, 2007). Ähnlich wie beim Lesen von Texten in anderen Fächern müssen in Mathematik Bedeutungen aus Symbolen erschlossen werden. Im Gegensatz zu anderen Fächern zeichnet sich Mathematik dadurch aus, dass Schüler nicht nur zwischen sprachlichen Ebenen, sondern auch zwischen weiteren Zeichensystemen (wie Graphen und Formeln) hin- und herwechseln können müssen (Adams, 2013; Prediger, 2013). Diese zusätzlichen bedeutungstragenden Systeme müssen bei der Konzeption von fachspezifischer Bildungssprache in Mathematik berücksichtigt werden (vgl. auch Duarte et al., 2011). Dementsprechend beschreibt Prediger (2013) bildungssprachliche Kompetenz in Mathematik als Wechseln zwischen verschiedenen, nicht vollständig deckungsgleichen Repräsentationen mathematischer Sachverhalten. Je nach Repräsentationsform werden unterschiedliche Aspekte mathematischer Beziehungen in den Vordergrund gestellt. Bildungssprachlich kompetente Schüler können je nach Situation die passende Repräsentation auswählen. Diese Entwicklungen deuten darauf hin, dass bei zukünftigen Untersuchungen der schwierigkeitsgenerierenden Merkmale von Mathematikaufgaben für Kinder mit nicht-deutscher Familiensprache auch fachliche und fachsprachliche Anforderungen der Aufgaben in den Blick genommen werden sollten.

4.4.2 Stellenwert bildungssprachlicher Anforderungen von Testaufgaben

Weitere an die vorliegende Arbeit anschließende Forschungsfragen beziehen sich auf das Zusammenspiel bildungssprachlicher Merkmale über die Schulzeit. Um das Konzept der Bildungssprache bezogen auf einzelne Klassenstufen besser greifbar zu machen, sollte zunächst genauer beschrieben werden, welche Merkmale bzw. Merkmalskombinationen typischerweise in verschiedenen Klassenstufen auftreten. Im deutschsprachigen Raum wurde in den letzten Jahren ein Literaturkorpus für Kinder im Grundschulalter erstellt, der allerdings vornehmlich auf außerschulischem Lesematerial basiert und sich nicht speziell auf bildungssprachliche Merkmale bezieht (childLex, Schroeder, Würzner, Heister, Geyken & Kliegl, im Druck). Anknüpfend an diesen Korpus könnten die sprachlichen Merkmale von schulischen Materialien wie beispielsweise Schulbüchern systematisch erfasst und über die Klassenstufen verglichen werden (vgl. Deane et al., 2006; Graesser et al., 2011 für entsprechende Analysen aus dem englischsprachigen Raum; Meuers, Berendes, Vajjala & Bryant, 2015 für erste Arbeiten aus dem deutschsprachigen Raum).

Außerdem sollte geprüft werden, ob sich der Einfluss bestimmter Merkmale auf das Lösen von Sachfachaufgaben über die Schulzeit hinweg verändert (vgl. Abschnitt 4.1). Sowohl in

der bisherigen Forschung als auch in der vorliegenden Arbeit konnte noch nicht geklärt werden, ob sich die Einflüsse einzelner bildungssprachlicher Merkmale von Testaufgaben in Sachfächern wie Mathematik oder Naturwissenschaften über Klassenstufen unterscheiden (vgl. Wolf & Leon, 2009). Erste Hinweise auf differenzielle Effekte stammen aus der bereits erwähnten Analyse von Mathematikitems für die Klassenstufen 4, 7 und 10, die von Shaftel et al. (2006) durchgeführt wurde. Der Einfluss sprachlicher Merkmale von Mathematikaufgaben auf die Schwierigkeit dieser Aufgaben nahm in dieser Studie über die Schulzeit hinweg tendenziell ab, was sich an einer geringeren Varianzaufklärung der Itemschwierigkeiten durch die sprachlichen Merkmale ablesen ließ. Um spezifische Effekte einzelner Merkmale zu prüfen, wäre ein über mehrere Klassenstufen verlinkter Aufgabenpool erforderlich. Idealerweise sollten zur Untersuchung dieser Fragestellung längsschnittliche Analysen durchgeführt werden. Neben korrelativen Studien zum Einfluss einzelner Merkmale könnten Unterschiede im Einfluss sprachlicher Anforderungen über Klassenstufen auch in experimentellen Studien geprüft werden. Die vorliegende Arbeit zeigte nur geringe Vorteile einer sprachlichen Vereinfachung von Mathematikaufgaben für Kinder mit nicht-deutscher Familiensprache, welche ein mittleres Sprachniveau erreicht hatten. Mit einer ähnlichen experimentellen Studie mit sprachlich erschwerten Aufgaben könnte geprüft werden, inwieweit sprachlich komplexere Aufgaben größere DIF-Effekte für Kinder mit nicht-deutscher Familiensprache nach sich ziehen könnten. Die Erschwerung könnte in einer solchen Studie daraus bestehen, dass den Aufgaben sprachliche Merkmale hinzugefügt würden, die erst für höhere Klassenstufen typisch sind.

4.4.3 Erfassung bildungssprachlicher Fähigkeiten

Die bisherige Forschung zu bildungssprachlichen Fähigkeiten bezieht sich—neben qualitativen Studien—zumeist auf Unterschiede in der Lösungswahrscheinlichkeit von Aufgaben mit unterschiedlich ausgeprägtem bildungssprachlichem Gehalt (Berendes et al., 2013). Viele Studien befassen sich demnach damit, wie sich der bildungssprachliche Gehalt von Texten bzw. Aufgaben messen lässt (Bailey et al., 2007; Deane et al., 2006; Graesser et al., 2011) und welche bildungssprachlichen Merkmale Zweitsprachlernenden besondere Schwierigkeiten bereiten (z. B. Abedi, 2009; Abedi & Lord, 2001; Duarte et al., 2011; Gogolin, 2009; Heppt et al., 2014; Shaftel et al., 2006; Townsend, Filippini, Collins & Biancarosa, 2012; Wolf & Leon, 2009). Hierbei wird häufig davon ausgegangen, dass zum Verständnis stärker bildungssprachlich geprägter Texte andere sprachliche Kompetenzen benötigt werden als zum Verständnis eher alltagssprachlicher Texte (vgl. Heppt, Henschel & Haag, eingereicht). Diese als komplexer angenommenen Schülerfähigkeiten werden jedoch nur selten direkt gemessen (vgl. Snow & Uccelli, 2009; Uccelli, Barr, et al., 2015; Uccelli,

Galloway, et al., 2015). Daher sollte empirisch geprüft werden, inwiefern sich diese Fähigkeiten von grundlegenderen sprachlichen Fähigkeiten unterscheiden lassen und ob sie engere Zusammenhänge mit den Leistungen in Sachfächern aufweisen als grundlegendere sprachliche Fähigkeiten.

Zum potenziellen Einsatz in Large-Scale Assessment-Studien eignen sich vor allem gruppenadministrierte Tests zu rezeptiven bildungssprachlichen Fähigkeiten. Fachübergreifende bildungssprachliche Schülerfähigkeiten werden im englischsprachigen Raum meist über das Verständnis von bildungssprachlichem Wortschatz erfasst (z. B. Townsend et al., 2012). Ein ergänzendes Instrument zur breiteren Erfassung bildungssprachlicher Fähigkeiten von Schülerinnen und Schülern (*Core Academic Language Skills*, CALS) wurde von Uccelli, Barr, et al. (2015) vorgeschlagen. Das Instrument umfasst sechs Teilbereiche bildungssprachlicher Fähigkeiten: Verständnis von verdichteten Satzstrukturen (Nominalisierungen, komplexe Syntax), Verständnis von Konnektoren und Diskursmarkern, Verständnis von anaphorischen Beziehungen (Wiederaufgreifen von bereits erwähnten Ideen), Fähigkeit zur argumentativen Textorganisation sowie Erkennen und Verwenden des bildungssprachlichen Registers (Uccelli, Barr, et al., 2015, S. 38f.). Die Autoren konnten nachweisen, dass es sich bei CALS um ein homogenes Konstrukt handelt. Schülerinnen und Schüler erwerben im Verlauf der Sekundarstufe zunehmend höhere CALS (Uccelli, Barr, et al., 2015), wobei die CALS von ELLs tendenziell geringer ausgeprägt sind. Ferner konnte gezeigt werden, dass CALS über die Effekte von ELL-Status, SES, Leseflüssigkeit und bildungssprachlichem Wortschatz hinaus einen eigenständigen Beitrag zur Varianzaufklärung von Lesefähigkeit leisten (Uccelli, Galloway, et al., 2015). Diese Befunde deuten darauf hin, dass CALS eine vielversprechende Operationalisierung bildungssprachlicher Fähigkeiten sein könnten. Auch aus dem deutschsprachigen Raum liegen gruppenadministrierte Testverfahren für bildungssprachliche Fähigkeiten von Grundschülerinnen und Grundschülern vor. Hierzu wurden Hörverstehenstests mit bildungssprachlich geprägten Hörtexten (Heppt et al., 2014), Tests für das Verständnis von Konnektoren (Heppt et al., 2012) sowie Tests für das Verständnis von Präfix- und Partikelverben (Uesseler et al., 2013) entwickelt. Für die Hörverstehenstests sowie für die Tests des Konnektorenverständnisses konnten jedoch keine Unterschiede zwischen Kindern deutscher Familiensprache und Kindern mit nicht-deutscher Familiensprache identifiziert werden (Heppt et al., 2012; Heppt et al., 2014). Entsprechende Messinstrumente für die Sekundarstufe wurden bislang im deutschsprachigen Raum nicht entwickelt. Die Entwicklung und Validierung von Tests zu bildungssprachlichen Fähigkeiten stellt daher einen wichtigen Bereich für zukünftige Forschung dar, um differenzielle Zusammen-

hänge zwischen bildungssprachlichen Fähigkeiten und der Leistung in Sachfächern empirisch besser überprüfbar zu machen.

5 Literatur

Aarts, R., Demir, S. & Vallen, T. (2011). Characteristics of academic language register occurring in caretaker-child interaction: Development and validation of a coding scheme. *Language Learning, 61*, 1173-1221.

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*, 231-257.

Abedi, J. (2008a). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*, 17-31.

Abedi, J. (2008b). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment, 13*, 193-214.

Abedi, J. (2009). Validity of assessments for English language learning students in a national/international context. *Estudios Sobre Educación, 16*, 167-183.

Abedi, J. & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*, 36-46.

Abedi, J., Hofstetter, C. H. & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*, 1-28.

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 219-234.

Abedi, J., Lord, C. & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Adams, T. L. (2013). Reading mathematics: More than words can say. *Reading Teacher, 56*, 786-795.

Aguirre-Muñoz, Z. (2000). *The impact of language proficiency on complex performance assessments: Examining linguistic accommodation strategies for English language learners.* (Doctoral dissertation, University of California at Los Angeles).

Anders, Y., Rossbach, H.-G., Weinert, S., Ebert, S., Kuger, S., Lehrl, S., et al. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*, 231-244.

Aukerman, M. (2007). A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction. *Reading Teacher, 60*, 626-635.

Avenia-Tapper, B. & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment, 20*, 95-111.

Bailey, A. L., Butler, F. A., LaFramenta, C. & Ong, C. (2004). *Towards the characterization of academic language in upper elementary science classrooms*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L., Butler, F. A., Stevens, R. & Lord, C. (2007). Further specifying the language demands of school. In A. L. Bailey (Hrsg.), *The language demands of school. Putting academic English to the test* (S. 103-156). New Haven, CT: Yale University Press.

Bailey, A. L. & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing, 28*, 343-365.

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In Deutsches Pisa-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323-407). Opladen: Leske + Budrich.

Baumert, J., Watermann, R. & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. Ein institutionelles und individuelles Mediationsmodell. *Zeitschrift für Erziehungswissenschaft, 6*, 46-72.

Berendes, K., Dragon, N., Weinert, S., Heppt, B. & Stanat, P. (2013). Hürde Bildungssprache? Eine Annäherung an das Konzept „Bildungssprache" unter Einbezug aktueller empirischer Forschungsergebnisse. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik: Interdisziplinäre Perspektiven* (S. 17-41). Münster: Waxmann.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L. & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores: A comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement, 72*, 754-773.

Böhme, K. & Weirich, S. (2012). Der Ländervergleich im Fach Deutsch. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 103-116). Münster: Waxmann.

Boonen, A. J. H., van Wesel, F., Jolles, J. & van der Schoot, M. (2014). The role of visual representation type, spatial ability, and reading comprehension in word problem solving: An item-level analysis in elementary school children. *International Journal of Educational Research, 68*, 15-26.

Bos, W., Tarelli, I., Bremerich-Vos, A. & Schwippert, K. (Hrsg.). (2012). *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Wendt, H., Köller, O. & Selter, C. (Hrsg.). (2012). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern im internationalen Vergleich*. Münster: Waxmann.

Bremerich-Vos, A. & Böhme, K. (2009). Lesekompetenzdiagnostik - die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 219-249). Weinheim: Beltz.

Brown, C. L. (2005). Equity of literacy-based math performance assessments for English language learners. *Bilingual Research Journal, 29*, 337-363.

Butler, F. A., Stevens, R. & Castellon, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. L. Bailey (Hrsg.), *The language demands of school. Putting academic English to the test* (S. 27-49). New Haven, CT: Yale University Press.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Hrsg.), *Educational Measurement* (4th ed., S. 220-256). Westport, CT: American Council on Education.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation, 19*, 104-120.

Carhill, A., Suarez-Orozco, C. & Paez, M. (2008). Explaining English language proficiency among adolescent immigrant students. *American Educational Research Journal, 45*, 1155-1179.

Chamot, A. U. & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.

Cole, N. S. & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

Cummins, D. D. (1991). Childrens interpretations of arithmetic word-problems. *Cognition and Instruction, 8*, 261-289.

Cummins, D. D., Kintsch, W., Reusser, K. & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19*, 121-129.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In N. H. Hornberger & B. Street (Hrsg.), *Encyclopedia of language and education, Volume 2: Literacy* (S. 71-83). New York: Springer.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1-28.

De Boeck, P. & Wilson, M. (Hrsg.). (2004). *Explanatory item response models. A generalized linear and nonlinear approach.* New York: Springer.

Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y. & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading, 10*, 257-275.

Demir-Vegter, S., Aarts, R. & Kurvers, J. (2014). Lexical richness in maternal input and vocabulary development of Turkish preschoolers in the Netherlands. *Journal of Psycholinguistic Research, 43*, 149-165.

Domenech, M. & Krah, A. (2014). What familial aspects matter? Investigating argumentative competences of learners at the beginning of secondary schooling in the light of family-based resources. *Learning, Culture and Social Interaction, 3*, 77-87.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Duarte, J., Gogolin, I. & Kaiser, G. (2011). Sprachlich bedingte Schwierigkeiten von mehrsprachigen Schülerinnen und Schülern bei Textaufgaben. In S. Prediger & E. Özdil (Hrsg.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit – Stand und Perspektiven der Forschung und Entwicklung in Deutschland* (S. 35-53). Münster: Waxmann.

Eckhardt, A. G. (2008). *Sprache als Barriere für den schulischen Erfolg. Potentielle Schwierigkeiten beim Erwerb schulbezogener Sprache für Kinder mit Migrationshintergrund*. Münster: Waxmann.

Ehlich, K. (1999). Alltägliche Wissenschaftssprache. *Info DaF, 1*, 3-24.

Ehlich, K. (2013). Sprachliche Basisqualifikationen, ihre Aneignung und die Schule. *Die Deutsche Schule, 105*, 199-209.

Ehmke, T. & Jude, N. (2010). Soziale Herkunft und Kompetenzerwerb. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt.* (S. 231-254). Münster: Waxmann.

Ercikan, K. & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Hrsg.), *Validity and test use: An international dialogue*

on educational assessment, accountability and equity (S. 69-86). Bingley, UK: Emerald Publishing.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D. & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education, 27*, 273-285.

Ercikan, K., Rubab, A., Law, D., Domene, J., Gagnon, F. & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*, 24-35.

Ernst-Slavit, G. & Mason, M. R. (2011). "Words that hold us up:" Teacher talk and academic language in five upper elementary classrooms. *Linguistics and Education, 22*, 430-440.

Faltis, C. J. (2013). Demystifying and questioning the power of academic language. In M. B. Arias & C. J. Faltis (Hrsg.), *Academic language in second language learning* (S. 3-26). Charlotte, NC: Information Age Publishing.

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L. & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*, 65-85.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Lambert, W., Stuebing, K. & Fletcher, J. M. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology, 100*, 30-47.

Gámez, P. B. & Lesaux, N. K. (2012). The relation between exposure to sophisticated and complex language and early-adolescent English-only and language minority learners' vocabulary. *Child Development, 83*, 1316-1331.

Gebhardt, M., Rauch, D., Mang, J., Sälzer, C. & Stanat, P. (2013). Mathematische Kompetenz von Schülerinnen und Schülern mit Zuwanderungshintergrund. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 275-308). Münster: Waxmann.

Gee, J. P. (2014). Decontextualized language: A problem, not a solution. *International Multilingual Research Journal, 8*, 9-23.

Genesee, F. & Lindholm-Leary, K. (2012). The education of English language learners. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major & H. L. Swanson (Hrsg.), *APA*

educational psychology handbook, Vol 3: Application to learning and teaching (S. 499-526). Washington, DC: American Psychological Association.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fähigkeiten. In I. Gogolin & U. Neumann (Hrsg.), *Streitfall Zweisprachigkeit – The Bilingualism Controversy* (S. 263-280). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gogolin, I. & Lange, I. (2011). Bildungssprache und Durchgängige Sprachbildung. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 107-127). Wiesbaden: VS Verlag für Sozialwissenschaften.

Graesser, A. C., McNamara, D. S. & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234.

Grießhaber, W. (2011). Zur Rolle der Sprache im zweitsprachlichen Mathematikunterricht. Ausgewählte Aspekte aus sprachwissenschaftlicher Sicht. In S. Prediger & E. Özdil (Hrsg.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit – Stand und Perspektiven der Forschung und Entwicklung in Deutschland* (S. 77-96). Münster: Waxmann.

Haag, N., Böhme, K. & Stanat, P. (2012). Zuwanderungsbezogene Disparitäten. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 209-235). Münster: Waxmann.

Hakuta, K., Butler, Y. G. & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara, CA: University of California Linguistic Minority Research Institute.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143-171). Berlin: Springer.

Hegarty, M., Mayer, R. E. & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology, 87*, 18-32.

Heinze, A., Herwartz-Emden, L., Braun, C. & Reiss, K. (2011). Die Rolle von Kenntnissen der Unterrichtssprache beim Mathematiklernen. Ergebnisse einer quantitativen Längsschnittstudie in der Grundschule. In S. Prediger & E. Özdil (Hrsg.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit – Stand und Perspektiven der Forschung und Entwicklung in Deutschland* (S. 11-34). Münster: Waxmann.

Heinze, A., Herwartz-Emden, L. & Reiss, K. (2007). Mathematikkenntnisse und sprachliche Kompetenz bei Kindern mit Migrationshintergrund zu Beginn der Grundschulzeit. *Zeitschrift für Pädagogik, 53*, 562-581.

Heppt, B., Dragon, N., Berendes, K., Stanat, P. & Weinert, S. (2012). Beherrschung von Bildungssprache bei Kindern im Grundschulalter. *Diskurs Kindheits- und Jugendforschung, 7*, 349-356.

Heppt, B., Henschel, S. & Haag, N. (eingereicht). Everyday and academic language proficiency: Investigating their relationships with school success and challenges for language minority learners.

Heppt, B., Stanat, P., Dragon, N., Berendes, K. & Weinert, S. (2014). Bildungssprachliche Anforderungen und Hörverstehen bei Kindern mit deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Pädagogische Psychologie, 28*, 139-149.

Hickendorff, M. (2013). The language factor in elementary mathematics assessments: Computational skills and applied problem solving in a multidimensional IRT framework. *Applied Measurement in Education, 26*, 253-278.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*, 1368-1378.

Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education, 16*, 159-188.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement, 74*, 627-658.

Huggins, A. C. & Elbaum, B. (2013). Test accommodations and equating invariance on a fifth-grade science exam. *Educational Assessment, 18*, 49-72.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J. & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*, 343-365.

Isaac, K. & Hochweber, J. (2011). Modellierung von Kompetenzen im Bereich „Sprache und Sprachgebrauch untersuchen“ mit schwierigkeitsbestimmenden Aufgabenmerkmalen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 186-199.

Johnson, E. & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*, 35-45.

Kane, M. (2010). Validity and fairness. *Language Testing, 27*, 177-182.

Kieffer, M. J., Rivera, M. & Francis, D. J. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments. 2012 update*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Kintsch, W. (1986). Learning from text. *Cognition and Instruction, 3*, 887-108.

Kintsch, W. & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92*, 109-129.

Kleemans, T., Segers, E. & Verhoeven, L. (2011). Cognitive and linguistic precursors to numeracy in kindergarten: Evidence from first and second language learners. *Learning and Individual Differences, 21*, 555-561.

KMK. (2004). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss der Kultusministerkonferenz von 15.10.2004*. München: Luchterhand.

Kölbl, C., Tiedemann, J. & Billmann-Mahecha, E. (2006). Die Bedeutung der Lesekompetenz für Sachfächer. *Psychologie in Erziehung und Unterricht, 53*, 201-212.

Kopriva, R. J. (2008). *Improving testing for English language learners*. New York: Routledge.

Kuhl, P., Harych, P. & Vogt, A. (2011). *VERA 3: Vergleichsarbeiten in der Jahrgangsstufe 3 im Schuljahr 2009/2010: Länderbericht Berlin*. Berlin: Institut für Schulqualität der Länder Berlin und Brandenburg.

Kuhl, P., Siegle, T. & Lenski, A. E. (2012). Soziale Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 275-296). Münster: Waxmann.

Lane, S. & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education, 39*, 165-214.

Lee, K., Ng, E. L. & Ng, S. F. (2009). The contributions of working memory and executive functioning to problem representation and solution generation in algebraic word problems. *Journal of Educational Psychology, 101*, 373-387.

Leiss, D., Schukajlow, S., Blum, W., Messner, R. & Pekrun, R. (2010). The role of the situation model in mathematical modelling—Task analyses, student competencies, and teacher interventions. *Journal für Mathematik-Didaktik, 31*, 119-141.

Leseman, P. P. M., Scheele, A. F., Mayo, A. Y. & Messer, M. H. (2007). Home literacy as a special language environment to prepare children for school. *Zeitschrift für Erziehungswissenschaft, 10*, 334-355.

Li, H. & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education, 25*, 327-346.

Linneweber-Lammerskitten, H. (2013). Sprachkompetenz als integrierter Bestandteil der mathematical literacy? In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Hrsg.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (S. 151-166). Münster: Waxmann.

Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437-448.

MacSwan, J. (2000). The threshold hypothesis, semilingualism, and other contributions to a deficit view of linguistic minorities. *Hispanic Journal of Behavioral Sciences, 22*, 3-45.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*, 333-368.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*, 160-179.

Mayer, R. E. & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg & T. Ben-Zeev (Hrsg.), *The nature of mathematical thinking* (S. 29-53). Hillsdale, NJ: Lawrence Erlbaum Associates.

McElvany, N., Becker, M. & Lüdtke, O. (2009). Die Bedeutung familiärer Merkmale für Lesekompetenz, Wortschatz, Lesemotivation und Leseverhalten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41*, 121-131.

McElvany, N. & Schwabe, F. (2013). Fairness von Lesetestaufgaben für Kinder aus Familien mit unterschiedlichem sozioökonomischem Status bei Large-Scale-Studien. In N. McElvany & H. G. Holtappels (Hrsg.), *Empirische Bildungsforschung - Theorien, Methoden, Befunde und Perspektiven. Festschrift für Wilfried Bos* (S. 219-234). Münster: Waxmann.

Messick, S. (1989). Validity. In R. Linn (Hrsg.), *Educational Measurement* (S. 13-103). Washington, DC: American Council on Education.

Meuers, D., Berendes, K., Vajjala, S. & Bryant, D. (2015). *Leseanforderungen in der Sekundarstufe: Ein Vergleich der linguistischen Komplexität von Schulbuchtexten.* Vortrag auf der 3. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Bochum.

Moreau, S. & Coquin-Viennot, D. (2003). Comprehension of arithmetic word problems by fifth-grade pupils: Representations and selection of information. *The British Journal of Educational Psychology, 73*, 109-121.

Müller, A. G. & Stanat, P. (2006). Schulischer Erfolg von Schülerinnen und Schülern mit Migrationshintergrund: Analysen zur Situation von Zuwanderern aus der ehemaligen Sowjetunion und aus der Türkei. In J. Baumert, P. Stanat & R. Watermann (Hrsg.), *Herkunftsbedingte Disparitäten im Bildungswesen. Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit vertiefende Analysen im Rahmen von PISA 2000.* (S. 221-255). Wiesbaden: VS Verlag für Sozialwissenschaften.

Müller, K. & Ehmke, T. (2013). Soziale Herkunft als Bedingung der Kompetenzentwicklung. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 245-271). Münster: Waxmann.

Mullis, I. V. S., Martin, M. O. & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In M. O. Martin & I. V. S. Mullis (Hrsg.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning* (S. 67-108). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nagy, W. & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*, 91-108.

Niklas, F. & Schneider, W. (2013). Home literacy environment and the beginning of reading and spelling. *Contemporary Educational Psychology, 38*, 40-50.

Noble, T., Rosebery, A., Suarez, C., Warren, B. & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*, 248-260.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B. & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching, 49*, 778-803.

Oliveri, M. E., Ercikan, K. & Zumbo, B. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education, 27*, 286-300.

Oliveri, M. E., Ercikan, K., Zumbo, B. D. & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—Comparing a latent class to a manifest DIF approach. *International Journal of Testing, 14*, 265-287.

Osterlind, S. J. & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks, CA: Sage.

Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1-4.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.

Pape, S. J. (2004). Middle school children's problem-solving behavior: A cognitive analysis from a reading comprehension perspective. *Journal for Research in Mathematics Education, 35*, 187-219.

Pennock-Roman, M. & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*, 10-28.

Pierce, M. E. & Melena, F. L. (2009). Designing vocabulary instruction in mathematics. *Reading Teacher, 63*, 239-243.

Pimperton, H. & Nation, K. (2010). Understanding words, understanding numbers: An exploration of the mathematical profiles of poor comprehenders. *The British Journal of Educational Psychology, 80*, 255-268.

Pöhlmann, C., Haag, N. & Stanat, P. (2013). Zuwanderungsbezogene Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 297-330). Münster: Waxmann.

Prediger, S. (2013). Darstellungen, Register und mentale Konstruktion von Bedeutungen und Beziehungen - mathematikspezifische sprachliche Herausforderungen identifizieren und bearbeiten. In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Hrsg.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (S. 167-183). Münster: Waxmann.

Prediger, S., Renk, N., Büchter, A., Gürsoy, E. & Benholz, C. (2013). Family background or language disadvantages? Factors for underachievement in high stakes tests. In A. M. Lindmeier & A. Heinze (Hrsg.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (S. 49-56). Kiel, Germany: PME.

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E. & Benholz, C. (2015). Sprachkompetenz und Mathematikleistung – Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen 10. *Journal für Mathematik-Didaktik, 36*, 77-104.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (Hrsg.). (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Reardon, S. F. & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853-891.

Reusser, K. (1997). Erwerb mathematischer Kompetenzen: Literaturüberblick. In E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 141-155). Weinheim: Beltz.

Richter, D., Kuhl, P. & Pant, H. A. (2012). Soziale Disparitäten. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 191-207). Münster: Waxmann.

Rindermann, H. & Baumeister, A. E. E. (2015). Parents' SES vs. parental educational behavior and children's development: A reanalysis of the Hart and Risley study. *Learning and Individual Differences, 37*, 133-138.

Rösch, H. & Paetsch, J. (2011). Sach- und Textaufgaben im Mathematikunterricht als Herausforderung für mehrsprachige Kinder. In S. Prediger & E. Özdil (Hrsg.), *Mathematiklernen unter Bedingungen der Mehrsprachigkeit – Stand und Perspektiven der Forschung und Entwicklung in Deutschland* (S. 55-76). Münster: Waxmann.

Sandilands, D., Oliveri, M. E., Zumbo, B. D. & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13*, 152-174.

Sato, E., Rabinowitz, S., Gallagher, C. & Huang, C.-W. (2010). *Accommodations for English language learner students: The effect of linguistic modification of math test item sets. (NCEE Report 2009-4079)*. National Center for Education Evaluation and Regional Assistance. Retrieved from http://www.eric.ed.gov/PDFS/ED510556.pdf.

Scarcella, R. (2003). Academic English: A conceptual Framework. Berkeley, CA: University of California Linguistic Minority Research Institute, UC Berkeley.

Scheele, A. F., Leseman, P. P. M. & Mayo, A. Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*, 117-140.

Scheele, A. F., Leseman, P. P. M., Mayo, A. Y. & Elbers, E. (2012). The relation of home language and literacy to three-year-old children's emergent academic language in narrative and instruction genres. *The Elementary School Journal, 112*, 419-444.

Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*, 431-459.

Schleppegrell, M. J. (2004). *The language of schooling. A functional linguistics perspective*. Mahwah, NJ: Erlbaum.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly, 23*, 139-159.

Schleppegrell, M. J. (2012a). Academic language in teaching and learning. *The Elementary School Journal, 112*, 409-418.

Schleppegrell, M. J. (2012b). Linguistic tools for exploring issues of equity. In B. Herbel-Eisenmann, J. Choppin, D. Wagner & D. Pimm (Hrsg.), *Equity in discourse for mathematics education: Theories, practices, and policies* (S. 109-124). Dordrecht: Springer Netherlands.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (im Druck). childLex - Eine lexikalische Datenbank zur Schriftsprache für Kinder im Deutschen. *Psychologische Rundschau*.

Schründer-Lenzen, A. (2008). Erklärungskonzepte migrationsbedingter Disparitäten der Bildungsbeteiligung. In J. Ramseger & M. Wagener (Hrsg.), *Chancenungleichheit in der Grundschule* (S. 107-116). Wiesbaden: VS Verlag für Sozialwissenschaften.

Schwabe, F. & Gebauer, M. M. (2013). (Test-)Fairness - eine Herausforderung an standardisierte Leistungstests. In N. McElvany, M. M. Gebauer, W. Bos & H. G. Holtappels (Hrsg.), *Jahrbuch der Schulentwicklung Band 17. Daten, Beispiele und Perspektiven. Sprachliche, kulturelle und soziale Heterogenität in der Schule als Herausforderung und Chance der Schulentwicklung* (S. 217-235). Weinheim: Beltz Juventa.

Schwabe, F., McElvany, N. & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50*, 219-232.

Schwabe, F., McElvany, N. & Trendtel, M. (im Druck). Reading skills of students in different school tracks: Systematic (dis)advantages based on item formats in large scale assessments. *Zeitschrift für Erziehungswissenschaft*.

Schwippert, K., Wendt, H. & Tarelli, I. (2012). Lesekompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 191-207). Münster: Waxmann.

Segeritz, M., Walter, O. & Stanat, P. (2010). Muster des schulischen Erfolgs von jugendlichen Migranten in Deutschland: Evidenz für segmentierte Assimilation? *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 62*, 113-138.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*, 105-126.

Sireci, S. G., Han, K. T. & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108-131.

Sireci, S. G. & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*, 170-187.

Sireci, S. G., Scarpati, S. E. & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490.

Slama, R. B. (2012). A longitudinal analysis of academic English proficiency outcomes for adolescent English language learners in the United States. *Journal of Educational Psychology, 104*, 265-285.

Slama, R. B. (2014). Investigating whether and when English learners are reclassified into mainstream classrooms in the United States: A discrete-time survival analysis. *American Educational Research Journal, 51*, 220-252.

Snow, C. E. & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Hrsg.), *The Cambridge handbook of literacy* (S. 112-133). New York: Cambridge University Press.

Solano-Flores, G. & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*, 3-13.

Spanos, G., Rhodes, N. C., Dale, T. C. & Crandall, J. (1988). Linguistic features of mathematical problem solving: Insights and applications. In R. R. Cocking & J. P. Mestre (Hrsg.), *Linguistic and cultural influences on learning mathematics* (S. 221-240). Hillsdale, NJ: Erlbaum.

Stanat, P. (2006). Disparitäten im schulischen Erfolg: Forschungsstand zur Rolle des Migrationshintergrunds. *Unterrichtswissenschaft, 34*, 98-124.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

Stanat, P., Rauch, D. & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider & P. Stanat (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt.* (S. 200-230). Münster: Waxmann.

Stern, E. (1992). Warum werden Kapitänsaufgaben „gelöst"? Das Verstehen von Textaufgaben aus psychologischer Sicht. *Der Mathematikunterricht, 38*, 7-29.

Stubbe, T. C., Tarelli, I. & Wendt, H. (2012). Soziale Disparitäten der Schülerleistungen in Mathematik und Naturwissenschaften. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 231-246). Münster: Waxmann.

Tarelli, I., Schwippert, K. & Stubbe, T. C. (2012). Mathematische und naturwissenschaftliche Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund. In W. Bos, H. Wendt, O. Köller & C. Selter (Hrsg.), *TIMSS 2011. Mathematische und*

naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (S. 247-268). Münster: Waxmann.

Tolar, T. D., Fuchs, L. S., Cirino, P. T., Fuchs, D., Hamlett, C. L. & Fletcher, J. M. (2012). Predicting development of mathematical word problem solving across the intermediate grades. *Journal of Educational Psychology, 104*, 1083-1093.

Townsend, D., Filippini, A., Collins, P. & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal, 112*, 497-518.

Uccelli, P., Barr, C. D., Dobbs, C. L., Galloway, P., Meneses, A. & Sanchez, E. (2015). Core Academic Language Skills (CALS): An expanded operational construct and a novel instrument to chart school-relevant language proficiency in pre-adolescent and adolescent learners. *Applied Psycholinguistics, 36*, 1077-1109.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A. & Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly, 50*, 337-356.

Uesseler, S., Runge, A. & Redder, A. (2013). „Bildungssprache“ diagnostizieren. Entwicklung eines Instruments zur Erfassung von bildungssprachlichen Fähigkeiten bei Viert- und Fünftklässlern. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik: Interdisziplinäre Perspektiven* (S. 42-67). Münster: Waxmann.

Ufer, S., Reiss, K. & Mehringer, V. (2013). Sprachstand, soziale Herkunft und Bilingualität: Effekte auf Facetten mathematischer Kompetenz. In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Hrsg.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (S. 185-201). Münster: Waxmann.

van den Noortgate, W. & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*, 443-464.

van Steensel, R. (2006). Relations between socio-cultural factors, the home literacy environment and children's literacy development in the first years of primary education. *Journal of Research in Reading, 29*, 367-382.

Verboom, L. (2008). Sprachbildung im Mathematikunterricht der Grundschule. In C. Bainski & M. Krüger-Potratz (Hrsg.), *Handbuch Sprachförderung* (S. 95-112). Essen: Neue Deutsche Schule Verlagsgesellschaft.

Vukovic, R. K. & Lesaux, N. K. (2013). The language of mathematics: Investigating the ways language counts for children's mathematical development. *Journal of Experimental Child Psychology, 115*, 227-244.

Walker, C. M., Zhang, B. & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*, 162-181.

Wendt, H., Stubbe, T. C. & Schwippert, K. (2012). Soziale Herkunft und Lesekompetenzen von Schülerinnen und Schülern. In W. Bos, I. Tarelli, A. Bremerich-Vos & K. Schwippert (Hrsg.), *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (S. 175-190). Münster: Waxmann.

Wolf, M. K. & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*, 139-159.

Wong Fillmore, L. & Snow, C. E. (2002). What teachers need to know about language. In C. T. Adger, C. E. Snow & D. Christian (Hrsg.), *What teachers need to know about language* (S. 7-53). McHenry, IL: ERIC Clearinghouse on Languages and Linguistics.

Wu, M. & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18*, 93-113.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*, 147-170.

Xie, Y. & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly, 50*, 403-416.

Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. *Educational Assessment, 14*, 122-138.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J. & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*, 170-192.

Young, J. W., King, T. C., Hauck, M. C., Ginsberg, M., Kotloff, L., Cabrera, J., et al. (2014). *Improving content assessment for English language learners: Studies of the linguistic modification of test items*. Princeton, NJ: Educational Testing Service.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

6 Anhang A: The Role of Academic-Language Features for Reading Comprehension of Language-Minority Students and Students From Low-SES Families

Dieser Beitrag ist in der Zeitschrift *Reading Research Quarterly* erschienen. Die Referenz lautet:

Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, *50*(1), 61–82. doi:10.1002/rrq.83

Der Link für den Download des Beitrags ist:
http://onlinelibrary.wiley.com/doi/10.1002/rrq.83/abstract

7 Anhang B: Second Language Learners' Performance in Mathematics: Disentangling the Effects of Academic Language Features

Dieser Beitrag ist in der Zeitschrift *Learning and Instruction* erschienen. Die Referenz lautet:

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24–34. doi:10.1016/j.learninstruc.2013.04.001

Der Link für den Download des Beitrags ist:
http://www.sciencedirect.com/science/article/pii/S0959475213000315

8 Anhang C: Linguistic Simplification of Mathematics Items: Effects for Language Minority Students in Germany

Dieser Beitrag ist in der Zeitschrift *European Journal of Psychology of Education* erschienen. Die Referenz lautet:

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: Effects for language minority students in Germany. *European Journal of Psychology of Education*, *30*(2), 145–167. doi:10.1007/s10212-014-0233-6

Der Link für den Download des Beitrags ist:
http://link.springer.com/article/10.1007%2Fs10212-014-0233-6

9 Anhang D: Effects of Mathematics Items' Language Demands for Language Minority Students: Do They Differ Between Grades?

Dieser Beitrag ist in der Zeitschrift *Learning and Individual Differences* erschienen. Die Referenz lautet:

Haag, N., Roppelt, A., & Heppt, B. (2015). Effects of mathematics items' language demands for language minority students: Do they differ between grades? *Learning and Individual Differences, 42,* 70-76*.* doi: 10.1016/j.lindif.2015.08.010

Der Link für den Download des Beitrags ist:
http://www.sciencedirect.com/science/article/pii/S1041608015001752